Visual attention modelling and applications. Towards perceptual-based editing methods

Olivier Le Meur

HABILITATION
Á DIRIGER DES RECHERCHES

présentée devant

L'université de Rennes 1
Spécialité : informatique

par

Olivier Le Meur

Visual attention modelling and applications. Towards perceptual-based editing methods

soutenue le 18/11/2014 devant le jury composé de :

Patrick Bouthémy,	Directeur de recherche INRIA, France	Président
Anne Guérin-Dugué,	Professeur, Univ. de Grenoble, France	Rapporteur
Philippe Salembier,	Professeur, Univ. de Catalogne, Espagne	Rapporteur
Laurent Itti,	Professeur, Univ. de Californie du Sud, USA	Rapporteur
Éric Marchand,	Professeur, Univ. de Rennes 1, France	Examinateur
Patrick Le Callet,	Professeur, Univ. de Nantes, France	Examinateur
Frédéric Dufaux,	Directeur de recherche CNRS, France	Examinateur

A Karen, Glen, Louen et Anouk.

Contents

Olivier Le Meur

Detailed Curriculum Vitae

Personal information

Date of birth: February 27, 1975
Nationality: French
Status: Married, 3 children

Professional information

Position: Associate Professor (HDR) in computer science and digital video at ESIR (Ecole Supérieure d'Ingénieurs de Rennes), `https://esir.univ-rennes1.fr/`

Teaching:
- Compression standards
- Image processing
- Visual attention understanding and modelling

Research interests:
- Visual perception and visual attention
- Image editing - inpainting, colorization, super-resolution
- Quality assessment
- Video compression

Address: IRISA, Campus de Beaulieu, 35042 Rennes Cedex - France
Mobile: +33 7 81 07 02 29
Phone: +33 2 99 84 74 25
Fax: +33 2 99 84 71 71
Email: olemeur@irisa.fr
Homepage: `people.irisa.fr/Olivier.Le_Meur/`
Google Scholar profile: `http://scholar.google.fr/citations?user=wl9Mj-8AAAAJ&hl=fr`

Education

2014 **Habilitation à Diriger des Recherches (HDR)**, *University of Rennes 1*, France.
Title: Visual attention modelling and applications. Towards perceptual-based editing methods

2002–2005 **PhD Thesis**, *University of Nantes*, France.
Supervisors: Prof. Dominique Barba and Prof. Patrick Le Callet.
Title: Computational modelling of visual attention and applications

1996–1999 **Master Degree in Computer Science**, *ENSSAT (Ecole Nationale Supérieure de Sciences Appliquées et de Technologie)*, Lannion, France.

Experience

2009–Present **Associate Professor in computer science and digital video**, *University of Rennes 1*, France, ESIR (Ecole Supérieure d'Ingénieurs de Rennes).

2005–2009 **Team leader at Technicolor R&D**.
The goal of the project I supervised concerned the modeling of the visual attention and its applications. Four peoples were involved in this project. The most important achievements concern on one hand the design of a computational model of visual attention and on other hand a saliency-based retargeting method. This method is currently embedded in a video compression scheme used for hand-held devices.

2002–2009 **PhD Thesis at Technicolor R&D**.

1999–2002 **R&D Engineer at Grass Valley**.
I was involved in the algorithm design for video compression scheme (MPEG-2 and MPEG-4).

Student supervision

PhD students

2014–2017 **Hristina Hristova**, *(50%)*, Hristina works on examplar-based style transfer for video sequences..

2012–2015 **Nicolas Dhollande**, *(70%)*, Nicolas works on the optimization of the new video compression standard HEVC in a context of ultra-high-definition. (Grant from Thomson Video Networks)..

2012–2015 **Julio Cesar Ferreira**, *(20%)*, Julio works on super-resolution algorithms for multi-view applications. [collaboration between Univ. Rennes 1 - Univ. Uberlandia Brazil, since Oct. 2013].

2011–2014 **Darya Khaustova**, *(80%)*, Darya investigates the quality assessment of 3D content by focusing more specifically on the visual deployment in 3D context. (Grant from Orange Labs).

2010–2014 **Mounira Ebdelli**, *(40%)*, Mounira works on spatio-temporal inpainting for loss concealment and video editing applications.

2009–2012 **Josselin Gautier**, *(50%)*, Josselin has worked on three aspects of the 3D workflow: compression of depth map, virtual view synthesis and visual attention in 3D context. Present-day position: post-doctoral fellowship at Anglia Ruskin university (UK).

2008–2012 **Brice Follet**, *(50%)*, Brice has worked on the modeling of visual attention by using top-down information. (Grant from Technicolor R&D).

2005–2008 **Alexandre Ninassi**, *(50%)*, Alexandre has worked on the use of saliency maps in the context of quality assessment. Intensive experiences have been conducted on video sequences both to assess their quality and to record eye movements. Present-day position: post-doctoral fellowship at ENSI CAEN. (Grant from Technicolor R&D).

Master students

2014 **Hristina Hristova**, *Handling aesthetics in globally illuminated scenes*.

2013 **Julien Sicre**, *Aesthetic prediction*.

2012 **Alan Bourasseau**, *Super-resolution algorithms*.

2011 **David Wolinsky**, *Virtual View synthesis by extrapolation*.

2010 **Younesse Andam**, *Saliency-based data-pruning*.

2008 **Clément Rousseau**, *Video retargeting*.

2007 **Ayodeji Aribuki**, *H.264 video compression*.

2006 **Guillaume Courtin**, *Video retargeting*.

2006 **Olivier Gaborieau**, *H.264 video compression*.

Participation of PhD defense committees

2015 **Darya Khaustova**, *Univ. of Rennes 1, France*, (co-supervisor).

2015 **Liu Yi**, *INSA, France*, (examinator).

2014 **Antoine Coutrot**, *Univ. of Grenoble, France*, (examinator).

2014 **Jiayu Liang Gary**, *Univ. of Hong Kong, China*, (examinator).

2014 **Hamed Rezazadegan Tavakoli**, *Univ. of Oulu, Finland*, (examinator).

2012 **Josselin Gautier**, *Univ. of Rennes 1, France*, (co-supervisor).

2012 **Brice Follet**, *Univ. of Paris VIII, France*, (co-supervisor).

2011 **Matthias Pizzoli**, *Univ. of Roma, Italia*, (examinator).

2010 **Matthieu Perreira Da Silva**, *Univ. of la Rochelle, France*, (examinator).

2009 **Alexandre Ninassi**, *Univ. of Nantes, France*, (co-supervisor).

2009 **Alexander Bur**, *Univ. of Neuchatel, Switzerland*, (examinator).

2007 **Mattei Mancas**, *Univ. of Mons, Belgium*, (examinator).

Scientific responsibilities

2012–2014 **Scientific supervisor**, *(grant PIIF-GA-2011-299202 SHIVPRO)*.
I hosts a foreign researcher (Prof. Z. Liu) for 2 (+1) years in the context of Marie-Curie project (Grant Agreement PIIF-GA-2011-299202 SHIVPRO). The topic of this project is about computational model of visual attention on high-resolution video sequences as well as salient region detection.

2012–2016 **Scientific excellence award (PES)**, *rating : A*.

2009–2013 **PERSEE**, *ANR Project*.
Perceptual coding for 2D and 3D contents. Person in charge : Vincent Ricordel (IRCCyN, Nantes). Collaborations : IRCCyN-Nantes, INRIA-Rennes, IETR-Rennes and le LTCI-TelecomParisTech.

2010–2013 **ARSSO**, *ANR Project*.
The ARSSO project focuses on multimedia content communication systems, characterized by more or less strict real-time communication constraints, within highly heterogeneous networks, and toward terminals potentially heterogeneous too. It follows that the transmission quality can largely differ in time and space. The solutions considered by the ARSSO project must therefore integrate robustness and dynamic adaptation mechanisms to cope with these features. Person in charge : Sylvaine Kerboeuf (Alcatel). Collaborations : INRIA-Grenoble-Rennes, CEA-LETI/LNCA, ALCATEL LUCENT BELL LABS, THALES Communications, EUTELSAT SA.

Referee IEEE Trans. On Image Processing, IEEE Trans. On CSVT, IEEE Trans. On Multimedia, Journal Of Vision, The Visual Computer, Cognitive Computation, Elsevier Image Communication, Journal of Selected Topics in Signal Processing, IEEE Transactions on Multimedia Computing Communications and Applications, ICIP, ICME, WIAMIS, EUSIPCO, QoMex, DICTA.

Scientific production

Publications 20 journals, 37 conferences and 2 book chapters. The complete list is given at the end of this CV.
Top 3 most cited papers (according to *Google Scholar* (November 19, 2014)):

- O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, *A coherent computational approach to model bottom-up visual attention*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28(5), pp. 802-817, 2006 (**cited by 368**);
- O. Le Meur, P. Le Callet, D. Barba, *Predicting visual fixations on video based on low-level visual features*, Vision research, Vol. 47(19), pp. 2483-2498, 2007 (**cited by 140**);
- A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, *Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric*, IEEE International Conference on Image Processing, pp. 169-172, 2007 (**cited by 108**).

Bibliometrics According to the *Google Scholar* web site, my H-index is equal to 18 (on November 19, 2014). **My number of citations is equal to 1355**.

Special sessions

- Technical Program Committee (TPC), DICTA 2014, QoMex 2014, EUVIP 2014, ICIP 2014, ICME 2014, ICME 2015.
- Co-organization (with Prof. Z. Liu (Shanghai University)) of a special session on visual attention at the conference, ICME 2014
- Co-organization (with Prof. Z. Liu (Shanghai University)) of a special session on visual attention at the conference WIAMIS 2013.
- Co-organization (with M. Mancas (Mons University)) of a special session on visual attention at the conference SPIE Photonics Europe 2012.
- Co-organization (with Prof. P. Le Callet (Nantes University)) of a special session on visual attention at the conference ICIP 2009, Egypte.

Special issue

- Special Issue in Signal Processing: Image Communication. Title: *Recent Advances in Saliency Models, Applications and Evaluations*. Guest editors: Zhi Liu, Olivier Le Meur, Ali Borji & Hongliang Li. Submission deadline December 2014.

Invited talks

- **UESTC** (University of Electronic Science and Technology of China), School of Electronic Engineering, Chengdu, China. Prof. Bing ZENG and Prof. Hongliang LI. Title of the talk: examplar-based inpainting methods, July 2014.
- **Sichuan University**, Chengdu, China. Prof. Qionghua WANG. Title of the talk: examplar-based inpainting methods, July 2014.
- National Research Group (**GdR**), Information Signal Image viSion (ISIS). Special session on visual attention. Title of the talk: Modelling the visual scanpath, 19^{th} June 2014.
- Invited plenary speaker at **KÉPAF 2013** (National conference of the Hungarian Association for Image Processing and Pattern Recognition).
- Presentation of my visual attention work at **LUTIN** (Cité de la science, Paris), 2013.

Languages

French **Mothertongue**

English **Fluent**

Publication list

Chapters

1. T. Baccino, O. Le Meur, B. Follet, **La vision ambiante-focale dans l'observation de scènes visuelles**, book chapter in *A perte de vue les nouveaux paradigmes du visuel, Presse du réel*, to be published in 2014.
2. L. Morin, O. Le Meur, C. Guillemot, V. Jantet, J. Gautier, **Synthèse de vues intermédiaires**, book chapter in, *Vidéo 3D: Capture, traitement et diffusion*, Lucas L., Loscos C. et Remion Y. (Editors), Hermès, 2013.

Journal papers [IF=Impact Factor]

1. Z. Liu, W. Zou and O. Le Meur, **Saliency Tree: A Novel Saliency Detection Framework**, IEEE Trans. On Image Processing, vol. 23, no. 5, pp. 1937-1952, May 2014 [IF=3.042].
2. Z. Liu, X. Zhang, S. Luo and O. Le Meur, **Superpixel-Based Spatiotemporal Saliency Detection**, IEEE Trans. on Circuits and Systems for Video Technology, vol. 24, no. 9, pp. 1522-1540, Sep. 2014 [IF=2.259].
3. Z. Liu, W. Zou, L. Li, L. Shen and O. Le Meur, **Co-saliency detection based on hierarchical segmentation**, IEEE Signal Processing Letters, Vol. 21(1), 2014 [IF=1.674].
4. C. Guillemot, O. Le Meur, **Image inpainting: overview and recent advances**, IEEE signal processing magazine, Vol. 31(1), pp. 127-144, January 2014. [IF=3.368].

5. C. Guillemot, M. Turkan, O. Le Meur and M. Ebdelli, **Object removal and loss concealment using neighbor embedding methods**, Elsevier Signal Processing: Image Communication, 2013.
6. O. Le Meur, M. Ebdelli and C. Guillemot, **Hierarchical super-resolution-based inpainting**, IEEE Trans. On Image Processing, Vol. 22(10), pp. 3779-3790, 2013. [IF=3.042].
7. Z. Liu, O. Le Meur, S. Luo and L. Shen, **Saliency detection using regional histograms**, Optics Letters, Vol. 38(5), March 2013 [IF=3.39].
8. O. Le Meur and T. Baccino, **Methods for comparing scanpaths and saliency maps: strengths and weaknesses**, Behavior Research Method, Vol. 45(1), pp. 251-266, March 2013. [IF=2.4]
9. J. Gautier and O. Le Meur, **A time-dependent saliency model mixing center and depth bias for 2D and 3D viewing conditions**, Cognitive Computation 2012.
10. B. Follet, O. Le Meur and T. Baccino, **New insights on ambient and focal visual fixations using an automatic classification algorithm**, i-Perception, Vol. 2(6), pp. 592-610, 2011.
11. F. Urban, B. Follet, C. Chamaret, O. Le Meur and T. Baccino, **Medium spatial frequencies, a strong predictor of salience**, Cognitive Computation, Vol.3, Issue 1, pp. 37-47, 2011 (Special issue on Saliency, Attention, Active Visual Search, and Picture Scanning).
12. B. Follet, O. Le Meur and T. Baccino, **Modeling visual attention on scenes**, Studia Informatica Universalis, Vol. 8, Issue 4, pp. 150-167, 2010
13. O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, **Do video coding impairments disturb the visual attention deployment?**, Elsevier, Signal Processing: Image Communication, Vol. 25, Issue 8, pp. 597-609, September 2010. [IF=0.836].
14. O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, **Overt visual attention for free-viewing and quality assessment tasks. Impact of the regions of interest on a video quality metric**, Elsevier, Signal Processing: Image Communication, vol. 25, Issue 7, pp. 547-558, 2010. [IF=0.836]
15. O. Le Meur and J.C. Chevet, **Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks**, IEEE Trans. On Image Processing, Vol. 19(11) pp. 2801-2813, 2010. [IF=3.315]
16. O. Le Meur, S. Cloarec and P. Guillotel, **Automatic content repurposing for Mobile applications**, SMPTE Motion Imaging journal, January/February 2010.
17. A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, **Considering temporal variations of spatial visual distortions in video quality assessment**, IEEE Signal Processing Special Issue On Visual Media Quality Assessment, 2009.
18. O. Le Meur, P. Le Callet and D. Barba, **Predicting visual fixations on video based on low-level visual features**, Vision Research, Vol. 47(19), pp 2483-2498, 2007. [IF=2.05]
19. O. Le Meur, P. Le Callet and D. Barba, **Construction d'images miniatures avec recadrage automatique basée sur un modèle perceptuel bio-inspiré**, Numéro spécial de la Revue Traitement du Signal (TS), Systèmes de traitement et d'analyse des Images, Vol.24(5), pp. 323-336, 2007.
20. O. Le Meur, P. Le Callet, D. Barba and D. Thoreau, **A coherent computational approach to model the bottom-up visual attention**, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 28(5), pp. 802-817, May 2006 [IF=4.3]

Conference papers

1. O. Le Meur and Z. Liu, **Saliency aggregation: Does unity make strength?**, ACCV, 2014.
2. L. Li, Z. Liu, W. Zou, X. Zhang and O. Le Meur, **Co-saliency detection based on region-level fusion and pixel-level refinement**, ICME, 2014.
3. N. Dhollande, O. Le Meur and C. Guillemot, **HEVC intra coding of Ultra HD video with reduced complexity**, ICIP, 2014.
4. J.C. Ferreira, O. Le Meur, E.A.B. da Silva, G.A. Carrijo and C. Guillemot, **Single image super-resolution using sparse representations with structure constraints**, ICIP, 2014.
5. D. Khaustova, J. Founier, E. Wyckens and O. Le Meur, **Investigation of visual attention priority in selection of objects with texture, crossed, and uncrossed disparities in 3D images**, HVEI, 2014.
6. D. Wolinski, O. Le Meur and J. Gautier, **3D view synthesis with inter-view consistency**, ACM Multimedia 2013.
7. Z. Liu and O. Le Meur, **Superpixel-based saliency detection**, WIAMIS 2013.
8. S. Luo, Z. Liu, L. Li, X. Zou and O. Le Meur, **Efficient saliency detection using regional color and**

spatial information, EUVIP, 2013.

9. M. Mancas and O. Le Meur, **Memorability of natural scenes: the role of attention**, ICIP 2013.
10. C. Guillemot, M. Turkan, O. Le Meur and M. Ebdelli, **Image inpainting using LLE-LDNR and linear subspace mappings**, ICASSP 2013.
11. M. Ebdelli, O. Le Meur and C. Guillemot, Image inpainting using LLE-LDNR and linear subspace mappings, ICASSP 2013.
12. D. Khaustova, , J. Fournier, E. Wyckens, O. Le Meur, **How visual attention is modified by disparities and textures changes**, SPIE HVEI 2013
13. O. Le Meur and C. Guillemot, **Super-resolution-based inpainting**, ECCV 2012. [Acceptance Rate=25
14. M. Ebdelli, O. Le Meur and C. Guillemot, **Loss Concealment Based on Video Inpainting for Robust Video Communication**, EUSIPCO 2012.
15. M. Ebdelli, C. Guillemot and O. Le Meur, **Examplar-based video inpainting with motion-compensated neighbor embedding**, ICIP 2012.
16. J. Gautier, O. Le Meur and C. Guillemot, **Efficient Depth Map Compression based on Lossless Edge Coding and Diffusion**, PCS 2012.
17. O. Le Meur, T. Baccino and A. Roumy, **Prediction of the Inter-Observer Visual Congruency (IOVC) and application to image ranking**, ACM Multimedia (long paper) 2011. [Acceptance Rate=17
18. B. Follet, O. Le Meur and T. Baccino, **Features of ambient and focal fixations on natural visual scenes**, ECEM 2011.
19. O. Le Meur, **Robustness and Repeatability of saliency models subjected to visual degradations**, ICIP 2011.
20. O. Le Meur, J. Gautier and C. Guillemot, **Examplar-based inpainting based on local geometry**, ICIP 2011.
21. J. Gautier, O. Le Meur and C. Guillemot, **Depth-based image completion for View Synthesis**, 3DTV Conf. 2011.
22. O. Le Meur, **Predicting saliency using two contextual priors: the dominant depth and the horizon line**, IEEE International Conference on Multimedia & Expo (ICME 2011) 2011 [Acceptance Rate=30
23. C. Chamaret, O. Le Meur and J.C. Chevet, **Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies**, ICIP, pp. 1077-1080, 2010 [Acceptance Rate=47
24. C. Chamaret, O. Le Meur, P. Guillotel and J.C. Chevet, **How to measure the relevance of a retargeting approach?**, Workshop Media retargeting, ECCV 2010.
25. C. Chamaret, S. Godeffroy, P. Lopez and O. Le Meur, **Adaptive 3D Rendering based on Region-of-Interest**, SPIE 2010.
26. O. Le Meur and P. Le Callet, **What we see is most likely to be what matters: visual attention and applications**, ICIP, pp. 3085-3088, 2009 [Acceptance Rate=45
27. C. Chamaret and O. Le Meur, **Attention-based video reframing: validation using eye-tracking**, ICPR, 2008.
28. A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, **Which Semi-Local Visual Masking Model For Wavelet Based Image Quality Metric?**, ICIP, 2008.
29. A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, **Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric**, ICIP, 2007.
30. A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, **Task impact on the visual attention in subjective image quality assessment**, EUSIPCO, 2006.
31. O. Le Meur, X. Castellan, P. Le Callet, D. Barba, **Efficient saliency-based repurposing method**, ICIP, 2006.
32. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, **A human visual model-based approach of the visual attention and performance evaluation**, SPIE HVEI, 2005.
33. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, **Modélisation spatio-temporelle de l'attention visuelle**, CORESA, 2005.
34. O. Le Meur, P. Le Callet, D. Barba, **Masking effect in visual attention modeling**, WIAMIS, 2004.
35. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, E. Francois, **From low level perception to high level perception, a coherent approach for visual attention modeling**, SPIE HVEI, 2004.

36. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, **Performance assessment of a visual attention system entirely based on a human vision modelling**, ICIP, 2004.
37. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, **Bottom-up visual attention modeling : quantitative comparison of predicted saliency maps with observers eye-tracking data**, ECVP, 2004.

Introduction

Over the last twenty years we have been experiencing an explosion of the number of captured images and videos. Social Web Communities such as Flickr, Youtube, Facebook and Instagram have collected a huge amount of photos and videos. Among the majors, Facebook is probably the largest. More than 250 billion photos have been uploaded on the social network site, and more than 350 million photos are uploaded every day on average [48]. At the end of 2011, the database of the company Instagram included over than 400 million photos with an upload rate of 60 images per second [78].

The exploitation of this massive amount of data users share requires advanced algorithms which could help organize and browse efficiently the contents. These problems fall with our domain of expertise. Recent work in this area concentrates on the link between computer vision and cognitive science. My research is part of this effort and focuses on the understanding and modelling of human visual attention and its applications in image editing.

This manuscript which constitutes a synthesis document of my research in preparation for my Habilitation degree (*Habilitation à Diriger des Recherches*) presents the most important outcomes of my research. Since my PhD degree in September 2005, I have been working on two main research themes which are the visual attention and saliency-based image editing. Before delving into the details of my research, a brief presentation of the visual attention and saliency-based image editing is made.

Visual attention: our visual environment contains much more information than we are able to perceive at once. To deal with this large amount of data, human beings have developed biological mechanisms and visual strategies to optimize the visual treatment. Out of those, the visual attention is probably the most important one. It allows to concentrate our biological resources over the most important parts of the visual field. Two kinds of visual attention have been identified: the covert and the overt visual attention. The former does not involve eye movements and refers to the act of mentally focusing on a particular area. The latter, involving eye movements, is used both to explore complex visual scenes and to direct the gaze towards interesting spatial locations. A number of studies [50] have shown that, in most circumstances,

overt shifts of attention are mainly associated with the execution of saccadic eye movements. Overt attention of attention is often compared to a windows to the mind. Saccade targeting is indeed influenced by top-down factors (the task at hand, behavioral goals, motivational state) and bottom-up factors (both the local and global spatial properties of the visual scene). The bottom-up mechanism, also called stimulus-driven selection, is the core of my research dealing with the visual attention. It occurs when a target item effortlessly attracts the gaze. My research consists in understanding and modelling this mechanism.

Saliency-based image editing: the high-level definition of image editing (as stated by Wikipedia) is the following: *image editing encompasses the processes of altering images*. These processes refer to color adjustments, histogram manipulation, noise reduction, inpainting, just to name a few. My research focusses on the use of the visual attention into image editing algorithms. As we will see, computational models of visual attention predict the most visually important areas within a scene. From an input picture, these models output a 2D saliency map which is a grey level map where the brighter areas indicate the highest saliency. Saliency-based image editing consists in altering images in function of the saliency map. To illustrate this general idea, an example is the retargeting approach (one of my former work which is not presented in this manuscript). The idea is to adapt automatically traditional contents to the specific constraints of small screen devices in order to provide users with the best possible viewing experience. The saliency map is in this case used to define the cropping window which should enclose as much as possible the saliency. Rather than displaying the whole content, only the content enclosed by the bounding box is displayed (for more details reader could refer to [110, 106]).

This documents is organized into four parts. The first three parts correspond to a research orientation. These parts are composed of several chapters which all include a short review of background material and our contributions in the given field. At the end of each chapter, we give the list of our scientific contributions. The link between these research orientations is made in the fourth part which presents my research perspectives. Note that, to ease the reading, parts as well as chapters are self-contained.

The first part is devoted to the theme of visual attention which was at the core of my Ph.D. work. This part is composed of 3 chapters dealing with the modelling of visual attention, eye tracking datasets and similarity metrics. Chapter 1 addresses the computational models of visual attention. A brief review of state-of-the-art models is first given before describing our main contributions. Chapter 2 presents methodological and practical guidelines for those who want to conduct an eye-tracking experiment. Parameters such as the cultural background, age, number of subjects, viewing duration, etc were discussed from the viewpoint of quality and bias effect for the eye tracking data. In addition, important points are emphasized such as the estimation of the inter-observers dispersion and center bias. In Chapter 3 we present a comprehensive review of

metrics used to evaluate the similarity degree between a ground truth (human saliency map or eye fixation data) and a prediction. Strength and limitations were discussed leading to some recommendations.

In the second part we report the research results on saliency-based applications. This part is composed of three chapters. In Chapter 4 we report the research results devoted to quality assessment. In fact, we investigated the link between quality assessment and visual attention. The underlying idea is to adjust the computation of quality score in function of the degree of interest represented by a saliency map. Extensive experiments have been done in order to establish the extent to which visual attention and quality are linked. Chapter 5 deals with a very recent research avenue regarding the prediction of how much an image is memorable. This research has been initiated by [79]. In [126], we performed eye tracking experiments to establish a link between visual attention and memorability. From the experiment outcome, we have proposed a new set of features to estimate the memorability of pictures. The last chapter 6 presents a machine learning application which predicts automatically the attractiveness score of an image. A model is trained by using eye data considering that the dispersion between observers is low when there is something in the picture that draws our attention. Thanks to this model, we can sort out a bunch of pictures according to their degree of attractiveness.

The third part addresses image inpainting. This part is composed of two chapters. Chapter 7 provides a comprehensive review of methods used to perform an examplar-based inpainting. Chapter 8 presents my most recent contribution in this research theme. The proposed method allows to deal with two limitations of examplar-based methods. The first one is the high sensitivity to parameters setting whereas the second is related to the fact that most of current methods are greedy.

At first glance, this research theme likely appears as rather 'distant' from the two previous themes. However methods and algorithms developed in this theme will be at the design basis of perceptual-based image editing methods as explained in the last chapter of this manuscript. The last chapter of the manuscript draws conclusions and provides new avenues for my research. They are grouped into 3 axes: visual attention, image editing and perceptually-based image editing.

Part I

Computational models of visual attention

Chapter 1

Computational models

1.1 Introduction

Computational saliency models are designed to predict where we look within a visual scene. Most of them are based on the assumption that there exists an unique saliency map in the brain. This saliency map, also called master map, aims at indicating where the most visually important areas are located. This is a comfortable view for computer scientist since the brain is compared to a computer as illustrated by figure 1.1. The inputs would come from our different senses whereas our knowledge would be stored in the memory. The output would be the saliency map which is used to guide the deployment of attention over the visual space.

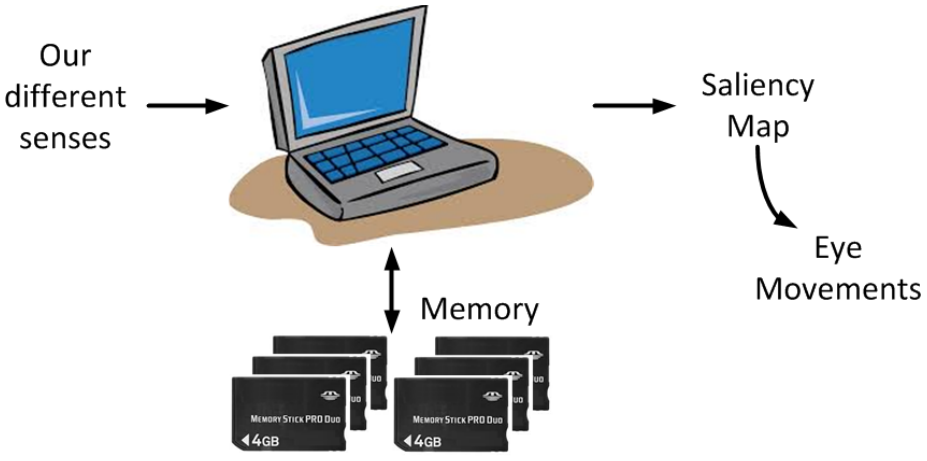


Figure 1.1: The brain as a computer, an unrealistic but convenient hypothesis.

From this assumption which is more than questionable, a number of saliency

models have been proposed. In section 1.2, we present a taxonomy and briefly describe the most influential computational models of visual attention. We will also describe our contributions in this field (section 1.3) which are related to robustness, the use of prior knowledge and dynamic saliency map for still pictures. The extension to video sequences is also discussed in section 1.4. We will conclude this chapter by emphasizing the strengths and limitations of current models.

1.2 Taxonomy

Since 1998, the year where the first computational and biologically plausible model of bottom-up visual attention was published by L. Itti, C. Koch and E. Niebur [84], there has been a growing interest on the subject. Indeed, several models, more or less biological and based on different mathematical tools, have been investigated. We proposed in 2009 a first saliency models taxonomy [108] which has been significantly improved and extended by Borji and Itti [13].
The taxonomy is composed of 8 categories as illustrated by figure 1.2 (extracted from [13]). A comprehensive description of these categories is given in [13]. Here we just give the main features of the four most important ones:

- **Cognitive models:** models belonging to this category rely on two seminal works: the Feature Integration Theory (FIT) [166] and a biological plausible architecture [97] for the computation of saliency map.
  The former relies on the fact that some visual features (called early visual features) are extracted automatically, unconsciously, effortlessly, and very early in the perceptual process. These features such as color, orientation, shape to name a few are automatically separated in parallel throughout the entire visual field. From the FIT, the first biological conceptual architecture has been proposed by Koch and Ullman [97]. This allows the computation of saliency map based on the assumption that there exists in the brain a single topographic saliency map. Models of this category follow a three-step approach: From an input picture, several early visual features are first extracted in a massively parallel manner, leading to one feature map per channel. A filtering operation is then applied on these maps in order to filter out most of the visually irrelevant information. Then, these maps are mixed together to form a saliency map.
  Some of our contributions are framed within this category, as pointed out by red arrows on figure 1.2.

- **Information theoretic models:** these models are grounded on a probabilistic approach. The assumption is that a rare event is more salient than a non rare event. The mathematical tool that can simply simulate this behaviour is the self-information. Self-information is a measure of the information amount carried out by an event. For a discrete random variable X, defined by $\mathcal{A} = \{x_1, ..., x_N\}$ and a probability density function, the amount of information of the event $X = x_i$ is given by

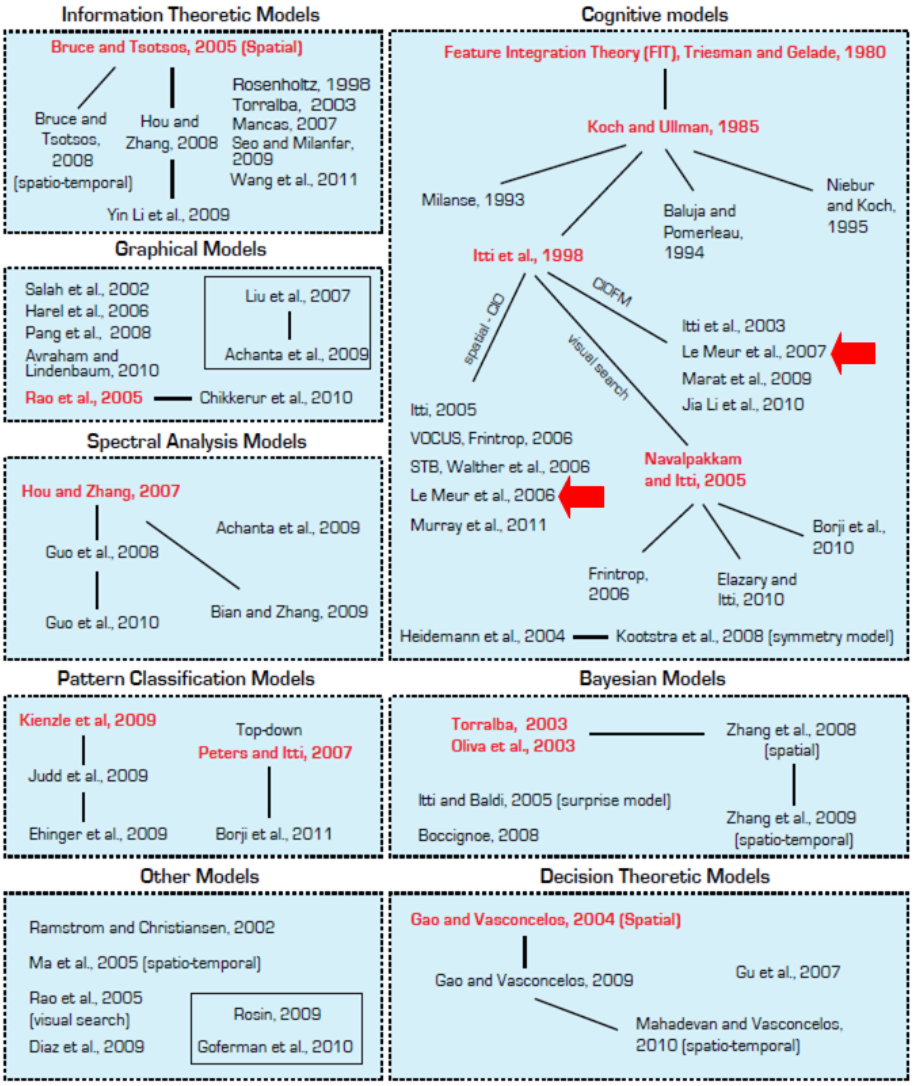


Figure 1.2: Taxonomy of computational model of visual attention. Courtesy of Borji and Itti [13].

$I(X = x_i) = -log_2\left(p(X = x_i)\right)$ bit/symbol.
The first model based on this approach has been proposed by Oliva et al. [137]. Bottom-up saliency is given by

$$S = \frac{1}{p(F|G)} \tag{1.1}$$

where, F denotes a vector of local visual features observed at a given

location while G represents the same visual features but computed over the whole image. When the probability to observe F given G is low, the saliency S tends to infinity. This approach has been re-used and adapted by a number of authors. The main modification is related to the support used to compute the probability density function:

  - Oliva et al. [137] determine the probability density function over the whole picture.
  - In [20] and [56], the saliency depends on the local neighbourhood from which the probability density function is estimated. The self-information [20] or the mutual information [56] between the probability density functions of the current location and its neighbourhood are used to deduce the saliency value.
  - A probability density function is learnt on a number of natural image patches. Features extracted at a given location are then compared to this prior knowledge in order to infer the saliency value [189].

- **Bayesian models:** the Bayesian framework is an elegant method to combine current sensory information and prior knowledge concerning the environment. The former is simply the bottom-up saliency which is directly computed from the low-level visual information whereas the latter is related to the visual inference, also called prior knowledge. This refers to the statistic of visual features in natural scene, its layout, the scene's category or its spectral signature to name a few. This prior knowledge which is shaped by our visual environment is one of the most important factors influencing our perception. It acts like a visual priming facilitating the scene perception and steering our gaze to specific parts.
  There exist a number of models using prior information, the most well known being the Theory of Surprise [81], Zhang's model [189], Oliva et al. [137].

- **Spectral analysis models:** This kind of model has been proposed in 2007 by Hou and Zhang [75]. The saliency is derived from the frequency domain based on the following assumption: *the statistical singularities in the spectrum may be responsible for anomalous regions in the image, where proto-objects are popped up.* From this assumption, they defined the spectral residual of an image which is the difference on a log amplitude scale between the amplitude spectrum of the image and its lowpass filtered version. This residual is considered as being the innovation of the image in the frequency domain. The saliency map in the spatial domain is obtained by applying the inverse Fourier transform. The whole process for an image

I is given below:

$$\begin{aligned}
\mathcal{A}(\mathbf{f}) &= \mathcal{R}\left(\mathcal{F}\left[I(\mathbf{x})\right]\right) && (1.2)\\
\mathcal{P}(\mathbf{f}) &= \phi\left(\mathcal{F}\left[I(\mathbf{x})\right]\right) && (1.3)\\
\mathcal{L}(\mathbf{f}) &= log\left(\mathcal{A}(\mathbf{f})\right) && (1.4)\\
\mathcal{E}(\mathbf{f}) &= \mathcal{L}(\mathbf{f}) - h(\mathbf{f}) * \mathcal{L}(\mathbf{f}) && (1.5)\\
\mathcal{S}(\mathbf{x}) &= g(\mathbf{x}) * \mathcal{F}^{-1}\left[exp\left(\mathcal{E}(\mathbf{f}) + \mathcal{P}(\mathbf{f})\right)\right]^2 && (1.6)
\end{aligned}$$

where, f is the radial frequency. $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the direct and inverse Fourier transform, respectively. $\mathcal{A}$ and $\mathcal{P}$ are the amplitude and phase spectrum obtained through $\mathcal{R}$ and ϕ respectively. h and g are two low-pass filters. This first approach has been further extended or modified by taking into account the phase spectrum instead of the amplitude one [65], quaternion representation and multiresolution approach [66].

1.3 Our cognitive model for still images

In 2006, we proposed a computational model of bottom-up visual attention [110]. The motivations were twofold.

The first one was simply to improve and to deal with the issues of the seminal work of L. Itti [84]. The most important drawback of Itti's model concerns the combination and the normalization of the feature maps which come from different modalities. In other words, the question is how to combine color, luminance and orientation information to get a saliency value. A simple and efficient method is to normalize all feature maps in the same dynamic range (e.g. between 0 and 255) and to sum them into the saliency map. Although efficient, this approach does not take into account the relative importance and the intrinsic features of one dimension compared to another.

The second motivation was our willingness to incorporate into the model important properties of the Human Visual System (HVS) which were rather neglected. These properties are related to the limited sensitivity of the HVS (we do not perceive all information present in the visual field with the same accuracy). They are simulated by Contrast Sensitivity Function and visual masking.

In the following sections, the global architecture of the proposed modelling is described as well as its main components. Its robustness to degraded pictures is presented. Then we will focus on two important improvements which are on one hand the use of high-level visual information and in the other hand the computation of a time-dependent saliency map.

1.3.1 Global architecture

Figure 1.3 illustrates the global architecture of the proposed model which is composed of two main parts, the visibility part and the saliency computation part. They are briefly described below. Readers could find more details in [110].

The visibility part aims to express visual information in terms of visibility threshold. The R, G and B components of the input picture are first transformed into an opponent-color space from which three components $\{A, Cr_1, Cr_2\}$ representing the achromatic, the blue-yellow and the red-green signals respectively are obtained. Contrast Sensitivity Functions (CSF) and visual masking are then applied in the frequency domain on the three components of the color space. The former normalizes the dynamic range of $\{A, Cr_1, Cr_2\}$ in terms of visibility threshold. Visual masking is then applied in order to take into account the influence of the spatial context on the visibility threshold. The visibility threshold of a given area tends to increase when its local neighbourhood is spatially complex. The 2D spatial frequency domain is then decomposed into a number of subbands which may be regarded as the neural image corresponding to a population of visual cells tuned to both a range of spatial frequency and orientation. These decompositions defined by psychophysics experiments leads to 17 channels for the achromatic component and only 5 channels for chromatic components.

Once the visual information has been coherently normalized, the second stage of the model consists in detecting the visually relevant parts of the image. Three operations are involved: chromatic-based reinforcement of the achromatic structures, center-surround filtering and perceptual grouping. The objective of these operations is summarized below:

- The chromatic-based reinforcement increases the magnitude of each site of the achromatic channels when the current site is surrounded by a high color contrast.
- The center-surround filter removes redundant information. The center-surround organization simulates the receptive fields of visual cells. These two regions provide an opposite response for the same stimulation.This filter is insensitive to uniform illumination and strongly respond on contrast. A difference of Gaussian, also called Mexican hat, is used to simulate the behaviour of visual cells.
- Perceptual grouping refers to the human visual ability to group and bind visual features to construct a meaningful higher-level structure. Here the perceptual grouping is a facilitative interaction based on the Gestalt principles of colinearity and proximity.

The filtered subbands are then combined into a unique saliency map. There exist a number of pooling strategies. In [25], seven feature combination strategies have been presented and evaluated. They are listed and briefly described below:

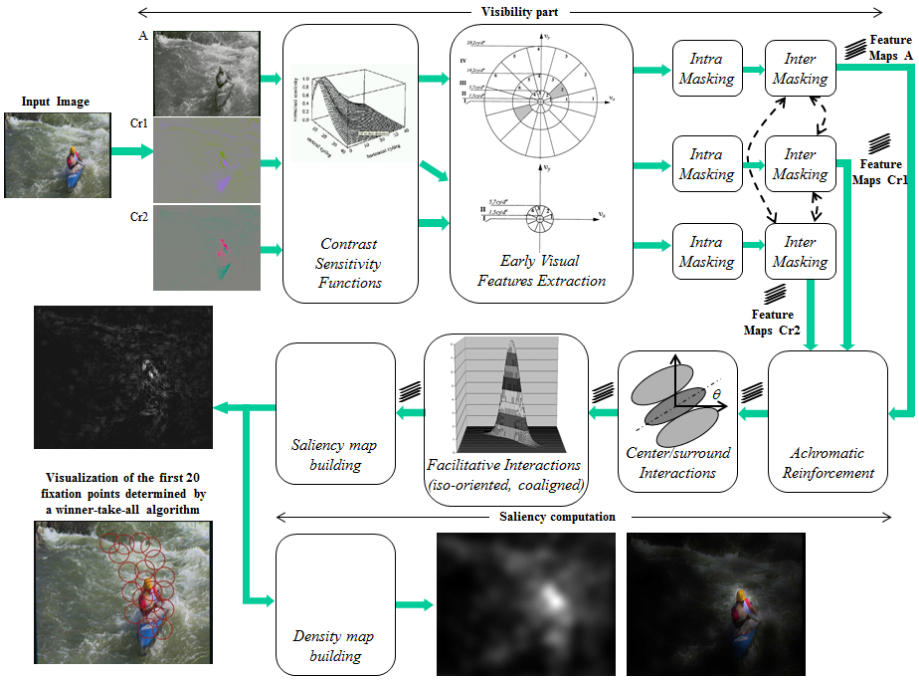


Figure 1.3: Architecture of the proposed model of bottom-up visual attention.

- NS, Normalization and Sum: This is the most simple method consisting in normalizing and summing all subbands into the final saliency map.
- NM, Normalization and Maximum: Compared to NS method, the summation is replaced by the maximum operator.
- CNS, Coherent Normalization and Sum: To normalize the subbands, the maximum saliency value for each visual dimension is empirically determined on a set of test pictures. These values are then used to perform the normalization.
- CNM, Coherent Normalization and Maximum: Compared to CNS method, the summation operator is replaced by the maximum operator.
- CNSP, Coherent Normalization, Sum plus Product: The idea here is to deal with the redundancies between feature maps. In other words, an item which would generate saliency in several visual dimensions should be more salient than an item generating saliency in only one dimension.
- CNMC, Coherent Normalization, intra and inter Map Competition: the CNSP approach is upgraded by using a WTA (Winner-Take-All) algorithm with localized inhibitory spread. The local maxima are then detected and

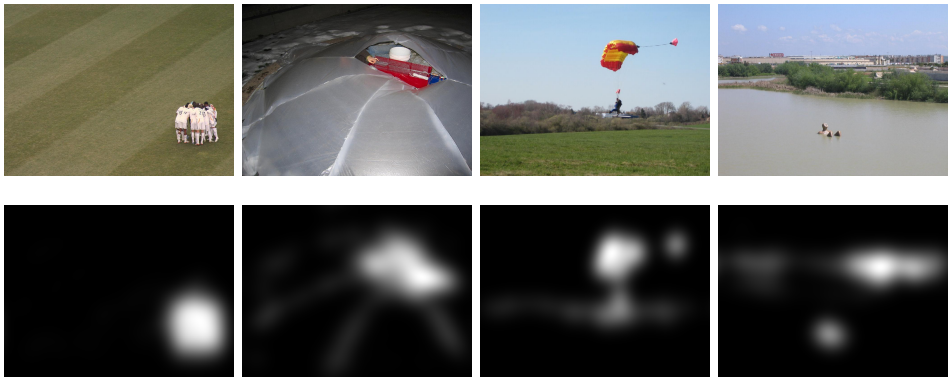


Figure 1.4: Examples of saliency maps predicted by the proposed model. Top: original images; Bottom: predicted saliency maps.

used to locally favour some parts of the picture. The number of maximum peaks, their values and the difference value between two consecutive maximum peaks are required to keep only the most interesting areas.

- GNLNS, Global Non-Linear Normalization followed by Normalization: This method implemented by L. Itti [83] consists in promoting the maps having few saliency peaks. This approach removes the maps having an uniform distribution or a high number of saliency peaks.

Figure 1.4 illustrates saliency maps computed by the proposed method (a simple fusion (NS) is here used). Saliency models perform well on this kind of images for which there is salient object on a simple background. Model performance significantly decreases in presence of high-level information [91] such as faces (*whether they are human or animal, real or cartoon, frontal or sideways, masked or not, in focus or blurry*), text (whether its font, size, quality) and horizon line which strongly attracts our attention [53]. Model performance is also low when the dispersion between observers is high (for more details, readers could refer to Chapter 2).
The natural step for improving the performance of saliency models is to take into account higher-level information. In the next section, we present an extension of our cognitive model.

1.3.2 Using prior knowledge: dominant depth and horizon line

In 2011, we improve the model performance [102] by inferring from low-level visual features two contextual information: the dominant depth and the position of the horizon line (if any).
Several studies support the hypothesis that there are separate neural pathways

for processing information about different visual properties [166]. These properties would be processed very quickly and unconsciously. The depth feature is one of them. We are indeed able to perceive the depth effortlessly. The most striking is that, even when we look at a picture, we are able to extrapolate the depth. As depth information is quickly available, this prior knowledge might affect eye movements. For instance, depth might contribute to an early recognition of the scene layout. In addition, from the knowledge of the dominant depth value, the average size of salient areas might be inferred. This property is used in the final pooling of the filtered subband. Given a dominant depth value, some subbands are favoured to get the final saliency map. The assumption is salient features are more likely to be present in low spatial frequencies for close-up scenes. For panoramic scenes, it might be more interesting to consider high frequencies than low spatial ones.
Foulsham et al. [53] provided evidence that the natural horizon line systematically attracts our visual attention. The position of the horizon line is then inferred from the low-level visual features to be used as a contextual prior. We simply propose to weight the final saliency map according to the spatial position of the horizon line.
We found [102] that the dominant depth does not bring a significant improvement when compared to a naive model. Regarding the horizon line, the median gain is of 2% in terms of AUC (Area Under Curve, see Chapter 3 for details).

1.3.3 Time-dependent model for static pictures

As previously mentioned, a number of computational models have been proposed to predict where we look. They all follow the same idea: a static picture is fed into the model which outputs a static saliency map. This map is supposed to represent the most visually salient parts of the scene. Although very easy and convenient to use, a static saliency map is not able to grasp the variety and complexity of visual guiding sources. Their influences can indeed increase or decrease over time. Moreover they are not necessarily concomitant but time-dependent. Some occurs after the stimulus onset, others appear after several second of viewing.
In 2012, we have designed a time-dependent saliency model [57] which outputs a video sequence of saliency maps from a static input picture. The problem is formalized as follow:

$$S(\mathbf{x}, t) = \sum_{k=1}^{K} p_k(t)\phi_k(\mathbf{x}) \tag{1.7}$$

where K is the number of visual guiding source, ϕ_k is a normalized 2D map representing the source k and $p_k(t)$ is the weight associated to the source k at a given time t. Compared to a traditional approach such as the naive summation (NS) presented in section 1.3.1, the difference lies in the presence of the time variable t.
In [57], five guiding sources have been considered:

- **Bottom-up saliency map**: this source represents the influence of low-level visual features on the gaze deployment. Several models (Itti [84], Bruce [20] and our model [110]) were used to compute this saliency map.

- **Center bias**: The strongest bias underlined by laboratory experiments is the central bias, also called re-centering bias. This is the tendency of observers to look at the screen's center whatever its interest (see chapter 2 section 2.3.2 for more information)

- **Foreground and background maps**: based on the fact that we are able to segment easily and quickly the figure from the ground, the depth map is split into two depth maps, one for the foreground and another for the background. To get these maps, the incoming depth map is thresholded at half the depth value through a sigmoid function, such that pixels values smaller and higher than 128 rapidly cancel out on background and foreground, respectively. Background values are modified such that the farther a point is in the background, the more it contributes to the background feature. At the opposite end, the closer a pixel is to the foreground, the more it contributes to foreground feature.

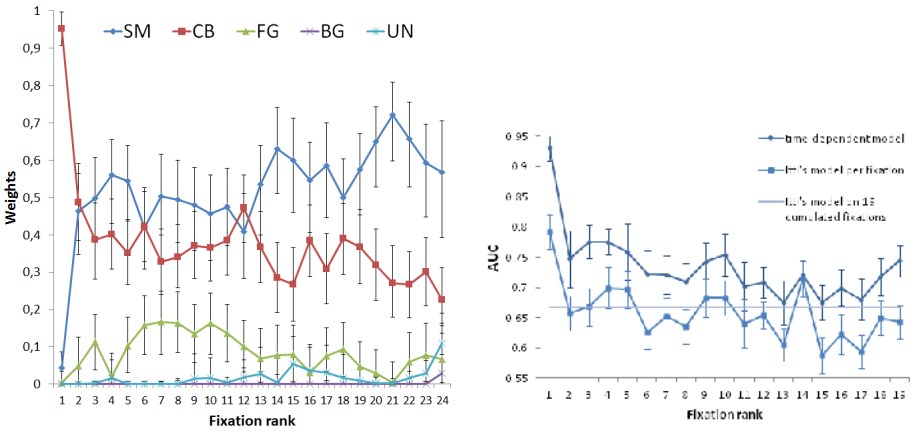
- **Uniform map**: to account for other guidance sources or top-down influences which could likely occur over time, an additional feature map, called uniform map, is used. For this map, all locations have the same probability to be fixated.

The weights $p_k(t)$ of the linear combination presented by equation 1.7 are inferred by an Expectation-Minimization algorithm. Figure 1.5 (a) gives the evolution of weights in function of the fixation rank. The most influent factor is the predicted low-level saliency. The central bias is strong and paramount on first fixation and decreases to a stable level from the third fixation. The foreground feature plays a non negligible role up to the 17^{th} fixations. At the opposite, the background feature is not a major guiding source of the visual attention. Finally, the contribution of the uniform distribution term remains low up to the late time of visualization. It models the influence of other high-level factors possibly due to top-down mechanisms that are not accounted by our proposed factors.

Once the weights have been estimated, the time-dependent saliency map is compared to Itti's model. Figure 1.5 (b) illustrates the results. The proposed model performs significantly better than Itti's model over time. The metric used for the evaluation is the AUC value (hit rate, see section 3.4 for more details).

1.3.4 Robustness

The invariance of saliency models subjected to degradation is important especially in the context of quality assessment and compression. The robustness of saliency models has been investigated in 2011 [103].
Different transformations and image processing filtering are applied to degrade



(a) (b)

Figure 1.5: (a) Temporal contributions (weights) of 5 features as a function of the fixation rank (SM: Salieny Map; CB: Center Bias: FG: Foreground; BG: Background; UN: Uniform). The error areas at 95% are computed by a bootstrap estimate (1,000 replications). (b) Evaluation of the performance with Itti's model.

the quality of a set of pictures. Figure 1.6 illustrates some of them applied on a given picture. The degradation operations are listed below:

- **Blur**: a Gaussian kernel of size 11×11 is used. Three values of variance are used: 1, 3 and 8. Obviously, the bigger the variance value the greater the smoothing produced.
- **Uniform variation of illumination**: the RGB components of the pictures are weighted by a fixed coefficient (0.2, 0.6, 1.4, 1.8). Coefficients greater than 1 tend to lighten the picture whereas a coefficient less than 1 darkens the picture.
- **Gaussian noise**: an independent Gaussian noise is added to the original image. The noise is with zero mean and a variance equal to 0.001, 0.01, 0.05 or 0.1. The bigger the variance the more the image is noisy.
- **Flip**: original pictures are flipped in right/left and up/down directions.
- **Rotation**: a rotation of the pictures is performed by an angle of 90, 180 and 270 (anti-clockwise) degrees. The rotation center is the picture's center. The invariance of models to rotation is interesting to investigate. Indeed Foulsham et al. recently provided evidences of a strong systematic tendency for saccades to occur along the axis of the natural horizon, whatever the picture orientation [53].

(a) Original (b) High blur (c) Under exposure (d) Over exposure

(e) Small noise (f) Strong noise (g) JPEG (Q=10) (h) JPEG (Q=1)

Figure 1.6: Examples of some degradations applied on picture (a).

- **JPEG coding**: a JPEG coding is applied on the original picture. The compression is performed by using the software *XnView*. Three quality factors (Q) are used: 40, 10 and 1. A small quality factor indicates a strong compression (or a low quality). For the smallest quality factor, strong block effects appear on the pictures, as illustrated by figure 1.6 (h).

A total of 2280 pictures (19 kinds of degradations multiplied by 120 pictures) is obtained. 2280×5 saliency maps plus those corresponding to original pictures have been computed, for five state-of-the-art models (Itti [84], Le Meur [110], Bruce [20], Judd [92], Hou [75]).

To measure the degree of similarity between saliency maps, a ROC (Receiver Operating Characteristic) analysis is performed (see section 3.3 for more details). Table 1.1 gives AUC values for each model and for the highest blur, noise, JPEG and luminance degradations. The average AUC value is very high for all models. It indicates that the predicted saliency maps are very similar (almost the same) whatever the visual degradations. It can be concluded that the repeatability of saliency models is very good. For the highest degradations, the lowest AUC value is equal to 0.82, that is still a good similarity indicator between predicted saliency maps (More results can be found in [103]).

1.4 Model for video sequences

To predict where observers look within a video sequence, it is necessary to consider a new feature which is the motion contrast. In this context, this is one of the most important visual attractors. With regards to dynamic complex scenes, previous studies such as [80] have indeed shown that the motion contrast is a much more reliable predictor of salient areas than other predictors such as

Table 1.1: Repeatability of saliency models for the highest degradations (average AUC values between predicted saliency maps (computed from original and impaired pictures)). Averages over degradations and models are given in the two last lines.

Model	Blur	Luminance		Noise	JPEG
	$\sigma^2 = 8$	Under	Over	$\sigma^2 = 0.1$	$Q = 1$
Itti	0.99	0.94	0.92	0.96	0.94
Le Meur	0.91	–	0.92	0.84	0.91
Bruce	0.97	0.99	0.82	0.95	0.95
Hou	0.99	0.99	0.88	0.88	0.96
Judd	0.97	0.88	0.92	0.86	0.90
Avg/Deg.	0.96	0.95	0.89	0.89	0.93
Avg/Model	0.95	0.89	0.93	0.94	0.90

luminance, orientation, etc. In this section, we present our contribution on this point. A survey of existing methods could be found in [13].

The basic aim of the temporal saliency map computation relies on the relative motion occurring in the retina. The relative motion is the difference between the local and the dominant motion. The local motion $\overrightarrow{V}_{local}$ at each point $\mathbf{s}$ of an image is the output of a hierarchical block matching. It is computed through a pyramid composed of a set of levels of different resolutions. For each level, the block matching is done for a certain neighbourhood size, that increases with the hierarchy level.

The local motion does not necessary reflect the motion contrast. This is only the case when the dominant motion is null, meaning that the camera is fixed. As soon as the camera follows something in the scene, it is necessary to estimate the global transformation that two successive images undergo. This global transformation, or the dominant motion, can be effectively estimated from the estimated local motion.

Assuming that the dominant motion is due to the camera, we estimate the global transformation between successive images I ($I : S \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3_+$) based on an estimated motion fields. The displacement $\overrightarrow{V}_\Theta(\mathbf{s})$, at a pixel site $\mathbf{s}$ ($\mathbf{s} \in \mathcal{S}$) related to a motion model parametrized by Θ is given by :

$$\overrightarrow{V}_\Theta(\mathbf{s}) = \begin{pmatrix} u_\Theta(\mathbf{s}) \\ v_\Theta(\mathbf{s}) \end{pmatrix} = B(\mathbf{s})\Theta \tag{1.8}$$

where $B(\mathbf{s})$ is the matrix of the parametric model used. $u_\Theta(\mathbf{s})$ and $v_\Theta(\mathbf{s})$ represent the horizontal and vertical displacement in regard to Θ parameters,

respectively. We consider the complete 2D affine motion model:

$$\overrightarrow{V}_{\Theta}(\mathbf{s}) = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \Theta \tag{1.9}$$

$$= \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} \tag{1.10}$$

where $\Theta = [a_1, a_2, a_3, a_4, a_5, a_6]^T$ are the affine parameters of the model. x and y represent the spatial coordinates of pixel $\mathbf{s} = (x, y)$. The six affine parameters are estimated with a popular robust technique based on the M-estimators [**?**]. To deal with the lack of stability of the standard least squares method to plausible outliers present in the data, the M-estimators lessen the outliers effects by replacing the squared residuals errors by a weighting function. The estimated affine parameters $\widehat{\Theta}$ have to minimize :

$$\widehat{\Theta} = arg \min_{\Theta} \sum_{\mathbf{s} \in S} \rho(r(\mathbf{s})) \tag{1.11}$$

where $r(\mathbf{s}) = I(\mathbf{s} + \overrightarrow{V}_{\Theta}(\mathbf{s}), t+1) - I(\mathbf{s}, t)$ represents the displaced frame difference (DFD). ρ is a symmetric, positive-definite function. We took here for ρ the Tukey's bi-weight function.
Once the six parameters have been computed, the relative motion $\overrightarrow{V}_{relative}$ representing the motion singularities is computed locally by subtracting the dominant motion from the local one. If the current location undergoes a dominant motion, the relative motion is null, otherwise the relative motion is non zero, indicating the presence of salient displacement. The relevance of the relative motion also depends on the average amount of the relative displacement over the picture. For example, a high relative motion is more conspicuous when there are only few relative displacements. To model such property, a linear quantification of $\overrightarrow{V}_{relative}$ is achieved in order to build a histogram. The median value of this histogram, called Γ_{median}, is a reliable estimator of the relative motion (note that this value is normalized according to the maximum velocity in order to be in the range 0 to 1. The maximum velocity is related to eye's ability to track an object). $\|\overrightarrow{V}_{relative}\|$ is then normalized according to Γ_{median}. Temporal saliency map S^T is finally given by:

$$S^T(\mathbf{s}) = \frac{\|\overrightarrow{V}_{relative}(\mathbf{s})\|}{\epsilon + \Gamma_{median}} \tag{1.12}$$

where ϵ is a small positive value to avoid the division by zero. The closer Γ_{median} to 0, the more the relative motion is perceptually important. This strategy simply reflects that it is easier to find a moving stimulus among stationary distractors (Γ_{median} close to 0) than a moving stimulus among moving distractors (Γ_{median} close to 1).

The spatial and the temporal saliency maps are then combined together in order to output the final spatio-temporal saliency maps. As discussed in

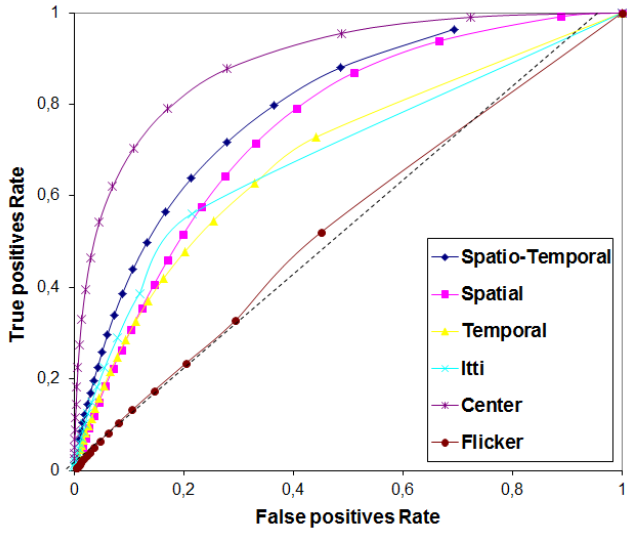


Figure 1.7: ROC curves for a set of video sequences. Four plausible models are compared (proposed Spatial, Temporal, Spatio-temporal and Itti's model). Two naive models which are the center and the flicker are also evaluated.

section 1.3.1, the pooling strategy is a difficult problem. We use the same strategies as those presented in section 1.3.1 to get the final result.

The performance of the proposed approach has been evaluated on a set of sequence (see [109] for more details). Figure 1.7 presents the ROC curve obtained by comparing two sequences of saliency maps. The proposed spatio-temporal model provides the best results compared to the spatial, temporal and Itti's model. Two naive models are also put to the test: the center model (see section 2.3.2 for more information) which provides the best results and the flicker model which is based on the frame difference of the input video. This model performs at the chance level.

1.5 Conclusion

During the last two decades the modelling and understanding of visual attention involving computer, neuroscience and cognitive scientists have made a lot of progress. The research is still stimulated by this cross-disciplinary collaboration. However, there are still a number of open issues to tackle. One of the most important concerns the prior knowledge of observers which is overlooked most of the time. The question is how can we introduce knowledge into a computational model of visual attention. Another issue is about the saliency models validation. There exist now a number of eye-tracking datasets which are more or less relevant or contaminated by experimental biases (see chapter 2 for a

review). The validation also requires to define tools to measure the similarity degree between ground truth and prediction. The ROC analysis is currently the most used metric although that this metric suffers from some important limitations.

1.6 Contribution in this field

There are several contributions in this field. The most important publications are listed below. They address the modelling of bottom-up visual attention on both still color pictures and video sequences. One dataset of eye-tracking data has been released to the community for research purposes.
In addition to the listed scientific contributions, I participated to the organization of four special sessions on viusal attention which took place during international conferences (ICIP 2009, SPIE 2012, WIAMIS 2013 and ICME 2014).

Journal:

- O. Le Meur, P. Le Callet, D. Barba and D. Thoreau, A coherent computational approach to model the bottom-up visual attention, IEEE Trans. on Pattern Analysis and Machine Intelligence, 28(5), May 2006.
- O. Le Meur, P. Le Callet and D. Barba, Predicting visual fixations on video based on low-level visual features, Vision Research, 47(19), pp. 2483-2498, Sept. 2007.
- O. Le Meur and J.C. Chevet, Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks, IEEE Trans. On Image Processing, 19(11), pp. 2801-2813, 2010.
- F. Urban, B. Follet, C. Chamaret, O. Le Meur and T. Baccino, Medium spatial frequencies, a strong predictor of salience, Cognitive Computation: 3(1), pp. 37-47, 2011 (Special issue on Saliency, Attention, Active Visual Search, and Picture Scanning).
- J. Gautier and O. Le Meur, A time-dependent saliency model mixing center and depth bias for 2D and 3D viewing conditions, Cognitive Computation, 3(2), pp. 141-156, 2012.
- Z. Liu, O. Le Meur, S. Luo and L. Shen, Saliency detection using regional histograms, Optics Letters, 38(5), 2013.
- Z. Liu, W. Zou, L. Li, L. Shen and O. Le Meur, Co-saliency detection based on hierarchical segmentation, IEEE Signal Processing Letters, vol. 21(1), 2014.

Conference:

- O. Le Meur and P. Le Callet, What we see is most likely to be what matters: visual attention and applications, ICIP, pp. 3085-3088, 2009.

- O. Le Meur, Robustness and repeatability of saliency models subjected to visual degradations, ICIP, pp. 3285-3288, 2011.

Chapter 2

Eye tracking datasets

2.1 Introduction

During the past decades, the rise of accessible commercial eye-tracking system has fuelled the visual perception research and strongly contributed to the emergence of new results and applications. Eye-tracking system is indeed a valuable tool to collect easily large amounts of data. These datasets serve different goals such as modelling, measuring, evaluating and understanding the way we look at visual scenes. More specifically this kind of datasets constituting a ground truth is used in the community to evaluate and compare the performance of computational models of visual attention. There exist more than a dozen datasets of eye tracking which can be freely downloaded from the Internet. However, there exist some limitations which are rather overlooked. Datasets are indeed intrinsically limited or bounded by different factors such as the number of observers, the viewing duration, the task at hand. More importantly datasets just provide a snapshot of the real world and can not grasp the complexity as well as the variety of our visual world.
The objective pursued here is to provide a comprehensive set of rules for releasing high quality eye tracking datasets. More specifically, these datasets are intended to serve as a ground truth for the evaluation and comparison of computational models of visual attention. We define the quality of an eye tracking dataset by two components, one related to the experimental methodology and the other related to the relevance of the tested materials. The former is simply about the most important factors which should be taken into account before starting the experiment. The latter concerns a post-processing phase which aims at detecting and filtering out images (stimuli) which do not bring any added values.
This chapter is organized as follows. In section 2.2, the main important methodological factors are elaborated. As surprising as it might seem, there is no standard describing accurately what must be done and what must be avoided to conduct an eye tracking experiment. In section 2.3, we present and evaluate

the main features of a set of popular datasets. Central bias, inter-observer dispersion, fixation duration and saccade lengths are examined. We will see that, despite the best efforts of their creators, there exist some strong biases which can strongly affect the experiment outcome. The aforementioned criteria can be advantageously used to detect and to discard dataset's outliers. In the last section, we will draw some conclusions.

2.2 Protocol

Conducting an eye tracking experiments is rather simple compared to other behavioural and psychophysics experiments. However, although simple, it is necessary to take care of some crucial parameters in order to guarantee that results are reproducible and comparable to existing ones. These parameters are listed and discussed below. Readers could also refer to [73] for complementary information.

2.2.1 Pixel Per Degree of visual angle

In vision research it is essential to specify the observation conditions as well as the stimuli. One important aspect for eye tracking is the visual angle subtended by stimuli and its angular resolution. The visual angle, expressed in degrees, represents the size of the stimulus on the retina. It is simply computed from the viewing distance and the size of the onscreen stimulus, as illustrated by figure 2.1. For a rectangular stimulus, the visual angle of its height θ_V (V stands for vertical) and the visual angle of its width θ_H (H stands for horizontal) should be given. They are obtained by

$$\theta_H = 2 \times arctan\left(\frac{H}{2d}\right) \tag{2.1}$$

$$\theta_W = 2 \times arctan\left(\frac{W}{2d}\right) \tag{2.2}$$

where H and V are the width and the height of the onscreen picture, respectively. It is important to mention that these sizes are not necessarily the screen's size. It depends on whether the stimulus covers the whole screen or not. d is the viewing distance, namely the distance between the eyes to the viewing plane.
The angular resolution of the stimulus expressed in pixels per degree (ppd) of visual angle is then deduced by dividing the stimulus's resolution by the visual angle. The angular resolution is used to normalize (or to be invariant to viewing conditions) indicators such as the saccade amplitudes, the fovea's size expressed in pixel, distance between two fixations points. H, V and d should be given in the same unit.

More importantly the angular resolution is essential to segment the raw eye tracking data into fixations and saccades. There are three standard measures to extract fixations from eye tracking data: dispersion, velocity and acceleration (for more details see [134],[156],[153]). Dispersion-based method relies on the

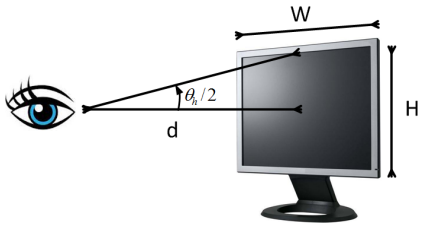
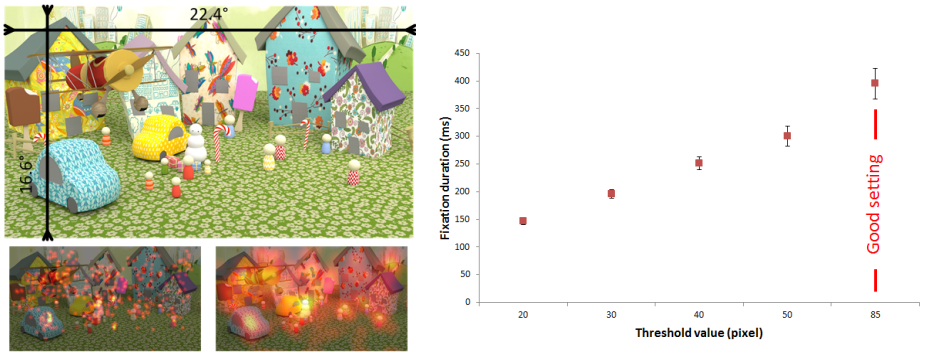


Figure 2.1: Visual angle θ_H subtended by stimuli. d is the viewing distance; H and W the height and width of the onscreen image, respectively.

distance between two consecutive samples. The distance is expressed in degree of visual angle. Two consecutive samples separated by a distance inferior to a user-predefined threshold are considered as belonging to the same visual fixation. Velocity and acceleration-based methods use a velocity or acceleration threshold, respectively, to identify samples related to saccade. The quality of these methods has been evaluated indicating that they perform quite well [10]. However, the major drawback of such method is the threshold value which is defined (or not) by the user [10], [156]. In the case of velocity-based method, if the threshold value is over-estimated, fixation durations might artificially increase. Under-estimation, in contrast, might cause a decrease of the fixation duration. In this case, as the mean fixation duration is an indicator of the cognitive load or the depth of the processing in the brain [172], the conclusion of the study might be erroneous. To illustrate this point, an eye tracking experiment, involving 18 observers, has been carried out. The picture used for the test is illustrated on figure 2.2 (a). Its resolution is of 1920×1080 pixels. The monitor (93×52cm) was positioned at a viewing distance of $2.34m$. The stimulus covering the whole screen subtended $22.4°$ horizontally and $16.6°$ vertically. A dispersion-based algorithm is used to identify fixations from the collected data. Figure 2.2 illustrates the protocol and gives the average fixation durations for different values of the threshold used by the dispersion-based algorithm. We just illustrate the influence of an under-estimation of the threshold value. The correct threshold value expressed in pixel per degree of visual angle is given by the horizontal resolution divided by the horizontal visual angle ($1920/22.4 = 85ppd$). When the threshold value is under-estimated, the average fixation duration decreases as illustrated by figure 2.2 (b). A low threshold value would suggest that observer is very close to the screen plane and that only a small portion of the picture falls within the fovea. This is the reason why the heat map (see section 2.3 to have more details about continuous saliency and heat maps) is composed of sparse and focussed points, as illustrated by bottom pictures of figure 2.2 (a).

2.2.2 Visual content

The visual content is at the heart of a vision experiment. In this section, we review the factors which should be taken into account when preparing the ex-



(a) (b)

Figure 2.2: (a) Top: visual stimulus and its visual angle. Bottom: heat map with a very low threshold value, equal to 20 ppd (leftmost value of graph (b)) and with the correct setting, 85 ppd (rightmost value of graph (b)); (b) Influence of the threshold value on the average fixation duration. Error bars correspond to the confidence interval. Courtesy of D. Khaustova.

periment.

Quality

Depending on the experiment purpose, the visual quality of the displayed stimuli might have a significant role. For instance, if the purpose of the experiment is to investigate the influence of a given factor on the fixation durations, it is important to collect stimuli of equivalent visual quality. Indeed Mannan et al.'s experiment [127] have shown that the fixation durations observed on blurred pictures are significantly higher than those obtained on unimpaired pictures. Judd et al. [89] reported observers make significantly fewer fixations on low resolution images than on high resolution image. One reason of the increase of fixation durations could be that, due to the presence of blur, observers need to dwell more on a particular location to accurately perceive it.
Regarding the localization of visual fixation, the stimulus's quality seems to have a low impact. In 1997, Mannan et al. [128] indeed showed that there is a high degree of similarity between fixations (in term of spatial coordinates) made by a group of observers to three versions of a given image (low-pass filtered, high-pass filtered, and unfiltered) (see also [89]). More recently, studies have shown that the linear correlation coefficient between saliency maps of unimpaired and impaired pictures is very high [101]. Figure 2.3 illustrates the distribution of fixation points for three levels of degradation of a given image. From the left-hand side to the right-hand side, the fixations have been collected when observers look at the unimpaired image, the image encoded by JPEG and by JPEG2000 respectively.

(a) (b) (c)

Figure 2.3: (a) Distribution of fixations collected when observers look at the unimpaired picture; (b-c) same as (a) but fixations were collected when observers look at the picture encoded by JPEG or JPEG2000, respectively. Yellow points correspond to visual fixations.

Image content

The image content itself may have a strong influence on our visual strategy. In next section we will see the inter-observer congruency (or the fixation dispersion) depends on the content. The factors influencing the congruency can stem from bottom-up (related to the stimulus) and top-down cues. For instances, human faces (as illustrated by figure 2.3), text, vanishing point, perspective, horizon line [53] have the capacity to attract our gaze. So, depending on the experiment's purpose, the selection of stimulus need to be done carefully.

2.2.3 With or without a task?

Low-level visual stimulus salience and the task at hand determine where people look in complex scenes [49]. Low-level visual salience refers to the ability of an area to attract our attention unconsciously and effortlessly. It relies on the low-level characteristic of visual stimuli, such as the color, the luminance, the texture, the motion [138, 161, 109] to name a few.
At the opposite, the visual deployment is task-dependent. The famous experiment of Yarbus in 1967 [186] is the perfect illustration of the influence of the

task on the visual deployment. Yarbus's contribution was to ask observer to view a painting seven times. For each viewing, a particular instruction is given before watching the image. Instructions were diverse such as *Estimate the material circumstances of the family in the picture*, *Give the ages of the people*, etc. Yarbus's conclusion was that the distribution of the points of fixations depends on the task in which the observer is engaged [186].
Depending on the purpose of the experiment, eye tracking can be carried out with or without instruction. It can be either a clear specification of a task to perform or just a simple instruction such as look at the picture. In this case, observers' eye movements are monitored while they freely viewed the stimuli. This is called a free-viewing task. The objective is here to favour the contribution of low-level visual salience in the eye movement guidance. However, even though participants were not given a task, it does not mean that top-down contributions do not exist. Top-down contributions are in fact due to the task at hand but also to the prior knowledge, motivations, mood and experience of observers. For the very first eye tracking experiment, task-based top-down contributions were supposed to be almost null when observers look at the stimuli freely or when a general instruction such as *look around* or *look at the picture* is given. However, even such simple instructions tell the subject something about what they are expected to do. In fact, the free viewing condition which is supposed to be void of top-down influences (except those related to observer's prior knowledge) is finally a kind of uncontrolled task. Given that observers are free to interpret the instruction on their own way, the variability of top-down effects can be high leading to a number of visual strategies, *with different observers employing different strategies and effectively performing different tasks* as stated by Tatler et al. [161]. To deal with this issue, Tatler et al. [161] instructed observers to perform a memory task. Two questions were asked to observers after each stimulus. This strategy aims at promoting the use of similar high-level mechanisms between participants, minimizing inter-observer dispersion. Apart from this important aspect this memory test is a way to motivate observers to pay attention and to keep focused on the experiment.
Another important top-down factor affecting eye movements is related to observer's prior knowledge. This point is addressed in the next section.

2.2.4 Observers

Observer's prior knowledge significantly influences our visual behavior. Despite the fact that this kind of contribution can not be ruled out, it seems important to underline and define the nature of this contribution. Prior knowledge is a general concept encompassing various aspects such as past experience, education, where we grow up or even social activities. This knowledge which is built progressively and continuously shapes our perception [132]. It then helps the visual system to better interpret the visual image and to resolve visual ambiguities. This prior knowledge about the scene, our familiarity with it have an influence, called contextual influence or scene-schema knowledge [69], on our visual deployment.

Cultural heritage

Regarding the cultural baggage, there is a current debate on whether it influences our visual strategy or not. In 2005, Chua et al. [28] demonstrated eye movement patterns differ between Chinese and Caucasian subjects when looking at complex scenes. Chinese participants tend to attend to the background information more than did American participants. However, in 2009, Evans et al. [47] cast doubt on previous results. They indeed did not find out difference in terms of eye movement patterns and recognition memory between American and Chinese participants, suggesting that both populations use the same strategies in scene perception. In 2011, Amatya et al. [3] examined the latency of reflexive saccades for a Chineses and Caucasian populations. Reflexive saccades are made in response to a sudden peripheral stimulus onset contrary to voluntary saccades which require additional cognitive processing [51]. Amatya et al. [3] observed that 29% of the Chinese subjects exhibited a high proportion of low latency saccades compared to those of Caucassian sujbects (only 2%). This study indicates that the cultural baggage would influence the visual strategy.

Age

Another aspect is related to observers' age. The older we are the more expert and familiar with our visual environment we are. So, given that prior knowledge depends on the age, the question is whether the age has an influence on the visual deployment or not. This question has been addressed by Acik et al. [1]. In this study they compared the viewing behaviour of three age groups: (1) 7 to 9 years old children; (2) 19 to 27 years old young adults; (3) a last group composed of people above 72 years. Participants have to perform a recognition task while they viewed natural and complex scenes. Results of this experiments showed that the low-level visual features are a more important source guidance for young children than adults. Authors suggest that the pregnancy of bottom-up features decreases with people's age. Beside, top-down influences which are related to prior knowledge might be more important for adults than young children in the guidance of eye movement. More recently, Musel et al. [129] showed that the nature of visual information extracted from the scene for a rapid categorization varies with age. Young participants tend to process a categorization task in a coarse-to-fine manner with a reaction time smaller than 100ms. For aged participants, the reaction time is longer indicating the presence of an additional visual process. Authors suggest that aged participants might use contextual information to correctly categorize visual scenes.

Number of observers

When the objective of the eye tracking experiment is to test an hypothesis, the number of subjects involved in the experiments is fundamental. Although difficult to define, the goal is to get sufficient sensitivity or statistical power in order to be able to draw significant conclusions. More specifically, the power analysis allows to calculate the minimum sample size required so that one can

be reasonably likely to detect an effect of a given size. The description of the power analysis is out of the scope of this chapter. Readers have to refer to Cohen's publications [29] to get a good overview and to Lenth's publication [116] in which some suggestions for successful and meaningful sample size determination are provided. In this section we propose to illustrate the power analysis by computing the number of observers required to obtain a statistically significant result.

Let's imagine we want to investigate the role of the global luminance of an image on the fixation duration. Two independent groups of observers participated to the experiment. Eye fixation will be collected when observers look at the original image and the modified image. How many observers should be involved in the experiment in order to show a significant effect on the average fixation duration? A t-test is used in order to evaluate whether the difference is statistically different or not. If one chooses the significance level ($\alpha = 0.05$) and the statistical power ($1 - \beta = 0.8$), the power of the t-test is given by $\delta = d\sqrt{N}$. δ is associated with the specific degree of power. To determine δ, you just need to read the power table (Table E.5 (Power as a function of δ and significance level α) in [76]). From this table δ must equal 2.8 for power equal to 0.8 and $\alpha = 0.05$. The d value (called the effect size) is given by $d = \frac{\mu_1 - \mu_2}{\sigma}$ and is used to test the hypothesis on the difference between the sample mean of average fixation durations μ_2 for the control and the tested group μ_1. σ is the standard deviation of the control population. N is the sample size of the control group. If you want to detect a difference of 10 ms between the two groups given a known standard deviation (here equal to 24), we have $d = 10/24 = 0.416$. Therefore the sample size should have a size of

$$n = \left(\frac{\delta}{d}\right)^2 \tag{2.3}$$

$$= \left(\frac{2.8}{0.416}\right)^2 \tag{2.4}$$

$$= 45.3 \tag{2.5}$$

Finally, if the experimenter wants to have a power of 0.8 to reject the equality between the two populations presenting a difference of 10 ms in terms of average duration of fixations, the number of participants involved in the experiment should be equal to 46.

Basic rules

Keeping in mind the aforementioned factors, it is also necessary to recall the basic rules used when preparing an experiment. Participants who are involved in an eye tracking experiment are most of the time naive to the purpose of the experiment and not familiar with this kind of experiment. They must have normal or corrected-to-normal vision. Note that there is no study, as far as we know, dealing with the influence of factors such as the socioeconomic status

and level of education on the visual strategy. It is however known that video-game playing enhances the capacity of visual attention [61, 38]. These studies reveal that video-game players responded more quickly to a stimulation than non video-game players and that with the same accuracy.

2.2.5 Viewing duration

Choosing the appropriate viewing duration is not so trivial. We provide hereafter some practical guidelines which could help you to take the right choice.
Several studies indicate that bottom-up influences were the highest for visual fixations that immediately followed the stimulus onset. The congruency between observers is then maximal (see next section for more details). In other words the consistency in fixation locations between observers decreases with prolonged viewing. Parkhurst et al. [138] explained this observation by the fact that the influence of bottom-up mechanisms decreases with the viewing time and is progressively overridden after several seconds of viewing by top-down mechanisms. Tatler et al. [161] revisited this hypothesis and conjectured that bottom-up mechanisms are not time-dependent and that low-level visual features might keep their ability to attract our visual attention throughout the viewing. They therefore explained that the decrease of consistency in fixation locations would be due to the growing influence of top-down mechanisms over time. Even if both studies propose different explanations, they all agree on the fact that bottom-up (or stimulus-dependent) mechanisms occur first. Based on the assumption that bottom-up influences vanished over time, studies investigating the influence of low-level visual salience endeavoured to design eye tracking experiments with rather short viewing time going most of the time from 2s to 10s. In Tatler et al. [161], the viewing time varied randomly between 1 and 10s to reduce predictability and training effect.

2.3 Comparative study of existing datasets

In this section we dress a list of existing databases of still images which are available on the Internet. Most of them are mainly used for evaluation and comparison of attention models. Table 2.1 gives the main characteristics such as viewing setup, participants. This table is extracted from [181], updated both with new links and from the paper [13]. Most important links to download these datasets can be found on http://stefan.winkler.net/resources.html.

2.3.1 Qualitative analysis

General view

From Table 2.1, some general remarks can be made. Most of datasets contain less than 200 hundred scenes, involve less than 30 observers and the viewing duration is smaller than 6 seconds. Some atypical points are noticeable: the dataset IRISA 2 involves 135 observers which are in fact split into 9 groups.

The dataset MIT LowRes is featured by 1544 stimuli which is in fact 8 groups of the same 193 images (for 8 spatial resolutions).
The number of pixel per degree of visual angle (nppd) is given in Table 2.1. It goes from 22 to 79 pixel per degree. The ppd value is here computed by dividing the angle subtended by the image's diagonal by its resolution. Most of the time authors specify the screen size but do not explicitly mention whether the stimuli were stretched to full screen or not. Another problem comes from the fact that the picture resolution varies. If the picture is not displayed in fullscreen mode, the nppd varies accordingly. Table 2.1 provides estimated ppd values based on the information provided by authors. Note that the sampling rate varies from 50 to 1000Hz.

Saliency map

A discrete fixation map noted f^i for the i^{th} observer is classically defined as below:

$$f^i(\mathbf{x}) = \sum_{k=1}^{M} \delta(\mathbf{x} - \mathbf{x}_{f^{(k)}}) \tag{2.6}$$

where $\mathbf{x}$ is a vector representing the spatial coordinates (x, y) and $\mathbf{x}_{f^{(k)}}$ is the spatial coordinates of the k^{th} visual fixation. The value M is the number of visual fixations for the i^{th} observer. $\delta(.)$ is the Kronecker symbol ($\delta(t) = 1$, if $t = 1$, otherwise $\delta(t) = 0$).
For N observers, the final fixation map f is given by:

$$f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f^i(\mathbf{x}) \tag{2.7}$$

A saliency map S is then deduced by convolving the fixation map by an isotropic bi-dimensional Gaussian function as described below:

$$S(\mathbf{x}) = f(\mathbf{x}) * G_\sigma(\mathbf{x}) \tag{2.8}$$

where σ is the standard deviation of the Gaussian. It is commonly accepted to use one degree of visual angle for σ. One degree of visual angle represents an estimate of the fovea's size. The standard deviation depends on the experimental setup (size of the screen, viewing distance, etc.). It is also implicitly assumed that a fixation can be approximated by a Gaussian distribution. An example of fixation and saliency maps is given on figure 2.4. A heat map (figure 2.4 (d)) which is a simple colored representation of the continuous saliency map (figure 2.4 (c)) is also shown. Red areas pertain for salient areas whereas the blue is used for the non salient areas. Note that the fixation map illustrated by figure 2.4 (b) is not stricto census the one defined by the formula (2.7).

Dataset	Scenes	Resolution	Users	Age	T [s]	D [cm]	d [in]	ppd	S	Eye Tracker	f [Hz]	Restraint
Toronto [19]	120	681×511	20		4	75	21	22	C			
IRCCyN 1 [110]	27	≈768×512	40		15			40	C	Cambridge Research	50	Chin rest
FiFA [24]	250	1024×768	7		2	80		–	C	EyeLink 1000	1000	Chin rest
Hwang [77]	160	1280×1024	30	19-40	10	–	19	46	C	EyeLink II	250	Chin rest
DOVES [171]	101	1024×768	29	μ =27	5	134	21	22	C	Fourward Tech. Gen. V	200	Bite bar
VAIQ [46]	42	varying	15	20-60	12	60	19	–	L	EyeTech TM3		
TUD Image 1 [123]	29	varying	20		10	70	19	–	C	iView X RED	50	Chin rest
MIT CSAIL [92]	1003	≈1024×768	15	18-35	3	61	19	42	L	ETL 400 ISCAN	240	Chin rest
MIT CVCL [42]	912	800×600	14	18-40		75	21	40	C	ISCAN RK-464	240	Head rest
Kienzle [96]	200	1024×768	14	–	3	60	19	29	C	–	–	–
GazeCom [37]	63	1280×720	11	18-34	2	45	22	63	C	EyeLink II	250	Chin rest
IRCCyN 2 [176]	80	481×321	18	19-45	15	40	17	–	L	Cambridge Research	50	
NUSEF [144]	758	1024×860	13	18-35	5	76	17	46	L	ASL	30	
Chikkerur [27]	220	640×480	8	18-35	5	70	–	–	L	ISCAN RK-464	240	Chin rest
MIT LowRes [89]	1544	1024×860	64	18-55	3	61	19	38		ETL 400 ISCAN	240	Chin rest
KTH [99]	99	1024×768	31	17-32	5	70	18	34	C	Eyelink I		Headmount
TUD Image 2 [2]	160	600×600	40		8	60	17	21	C	iView X RED	50	Head rest
TUD Interactions [146]	54	768×512	14	22-35		70	17	–	C	SMI	50/60	Chin rest
MIT Benchmark [90]	300	≈1024×768	39	18-50	3	61	19	30		ETL 400 ISCAN	240	Chin rest
McGill ImgSal [121]	235	640×480	21			70	17	24	L	Tobii T60	60	
IRISA 1 [126]	15×9	384×384	17	22-37	5	65	19	25	L	Face Lab 5	60	
IRISA 2 [94]	6(× 9)	1920×1080	135	21-60	20	60	85	79	L	Tobii X50	50	Chin rest

Table 2.1: Eye tracking datasets at a glance (T is viewing time, D is viewing distance, d is screen diagonal, f is the sampling rate ot the eye tracker). Screen(S)=[C=CRT; L=LCD]. Adapted from [181].

(a) Original (b) Fixation map (c) Saliency map (d) Heat map

Figure 2.4: Example of fixation (b), saliency (c) and heat map (d). The red dots on (b) represent the fixation points.

2.3.2 Quantitative analysis

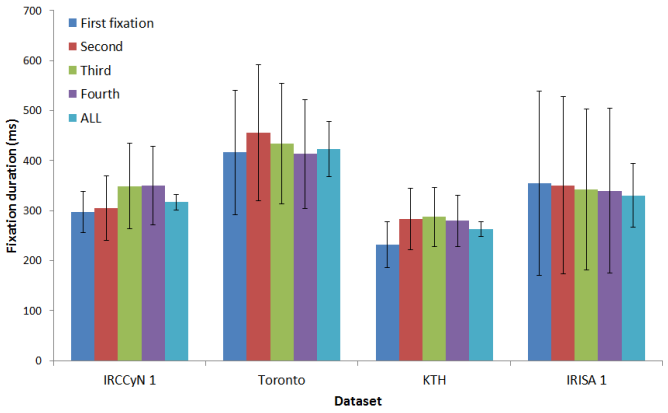
A quantitative analysis is performed on four datasets which are commonly used as a ground truth for evaluating the performance of visual attention models. This analysis is carried out on the basis of two parts: the first one is related to the main components of the visual scanpath which are the fixation and saccade. The second part is composed of the central bias and the inter-observer congruency. These two indicators are strongly correlated to the visual scene and would indicate whether a saliency model would be able to predict the salient areas easily or not.

Fixation durations

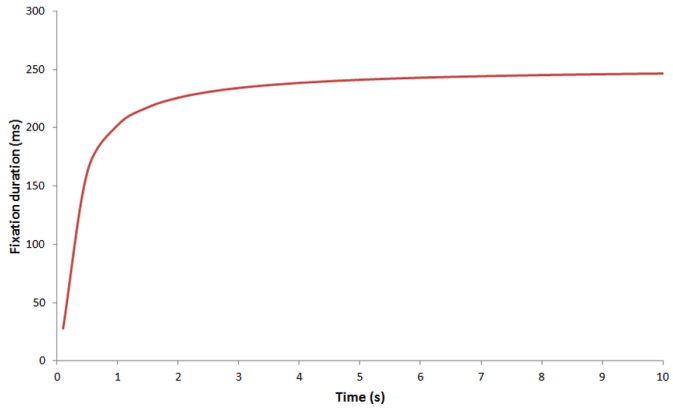
Figure 2.5 gives the average duration of the first, second, third, fourth and all visual fixations for four state-of-the-art datasets. The average fixation duration varies between 200 and 600 ms. For IRCCyN 1 and KTH, the average tends to increase over time. For Toronto and IRISA 1, we do not retrieve this trend. The duration of visual fixation is in fact dependent on a number of factors. First let's start by pointing out that the duration of the visual fixation is often considered as reflecting the deep and the speed of the visual processing in the brain. The longer fixation duration is, the deeper the visual processing is [70]. Total fixation time (i.e. cumulative duration of fixations within a region) can be used to gauge the amount of total cognitive processing engaged with the fixated information [145].
Overall, it is well admitted that the fixation duration increases over time to reach an asymptote. Unema et al. [169] proposed a parametric model to predict the fixation duration in function of the viewing time. This model is given by:

$$fd(t) = 252 \times exp\left(-\frac{0.22}{t}\right) \tag{2.9}$$

where t is the time at a resolution of 0.5s. The value 252 is the asymptotic value where 0.22 is called the acceleration rate. These values have been estimated in a particular context. This is illustrated by figure 2.5 (b).



(a)



(b)

Figure 2.5: (a) Average fixation durations for four state-of-the-art datasets. Error bars correspond to the confidence interval. (b) Parametric model of fixation durations in function of the viewing time.

Saccade amplitude

Figure 2.6 (a) illustrates the average amplitude saccades for the four datasets. The first four average amplitudes of saccades are given. The average saccade amplitude increases over time to reach an asymptote.
There is a considerable difference in average saccade amplitudes between the couples IRCCyn 1 / KTH and IRISA / Toronto datasets. One reason may be due to the visual content. We know that the number of objects in the scene affects the average saccade amplitude. In [169], the average saccade size was 6.3 degrees in the scenes containing 8 objects and 5.6 in the scenes containing 16 objects. Another reason is about the central bias which is more or less pronounced for

these datasets (see section *Central bias* for more details). Finally, the viewing condition and more specifically the ppd value plays an important role. It is interesting to note that the ppd values for IRCCyN 1 and KTH is equal to 40 and 34 respectively whereas the ppd values for IRISA 1 and Toronto is 25 and 22 respectively. Figure 2.6 (b) gives the saccades distribution of the four same datasets. Distributions is positively skewed with a mode between one and two degrees. The mean of saccades amplitudes is given in Figure 2.6 (a). Saccade amplitude distributions of IRCCyN 1, KTH and IRISA datasets follow a Gamma distribution [72]. The Gamma distribution starts at the origin and has a shape which is parametrized by two positive-defined parameters. The probability density function of the Gamma distribution is given by

$$y = \frac{1}{\theta^k \Gamma(k)} x^{k-1} exp\left(-\frac{x}{\theta}\right) \tag{2.10}$$

where, y and x represent the probability and the saccade amplitude respectively. k and θ are the shape and the scale parameters of the distribution, respectively. $\Gamma(.)$ is the Gamma function.
Parameters k and θ of the Gamma distribution are estimated by matching moments (chapter 22 of [52]). The scale and shape parameters are given by $\theta = \frac{s^2}{\overline{x}}$ and $k = \left(\frac{\overline{x}}{s}\right)^2$. $\overline{x}$ and s are the sample mean and the sample variance respectively. There are equals to $[1.839, 1.416]$, $[1.996, 3.01]$, $[2.296, 1.293]$, $[2.04, 2.17]$ for the IRCCyN 1, Toronta, KTH and IRISA 1 datasets. On figure 2.6, there is a slight difference between the probability density function of saccade amplitude and the Gamma distribution due to the quantization applied on raw data (the bins size of the histogram of Figure 2.6 (b) is equal to one degree). The two parameters of the Gamma distribution vary from one dataset to another. This is obviously due to the experiment setup (viewing time, viewing distance). For instance, the influence of viewing time on the scale and shape parameters is assessed on IRCCyN's dataset. Table 2.2 gives the k and θ values in function of the fixation rank. The observed trend is a decrease of the k value and an increase of θ value. On the right-hand side of table 2.2, the corresponding Gamma distributions are plotted. Results indicate that the pdf's mode is equal to 0.35 for the first fixation and decreases to reach 0.28 for the fourth fixation. It would suggest that the proportion of rather small saccade amplitudes decreased over the viewing time. One plausible reason is that observers might first focus on the salient region just after the stimulus onset and then start to explore more intensively the scene. The observed results can also be related to the inter-observer congruency. As explained in the next section, the inter-observer congruency decreases with the viewing time.

Inter-observer congruency

The inter-observer congruency allows to evaluate the level of visual agreement of the visual strategy employed by observers. Two methods scoring the visual agreement between 0 and 1 can be used: a one-against-all approach (also called

	k	θ
1	2.13	1.04
2	2.25	0.98
3	2.18	1.16
4	2.07	1.25
ALL	1.84	1.41

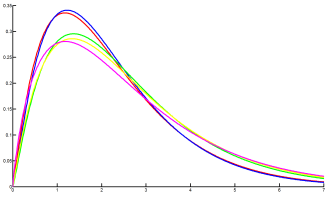
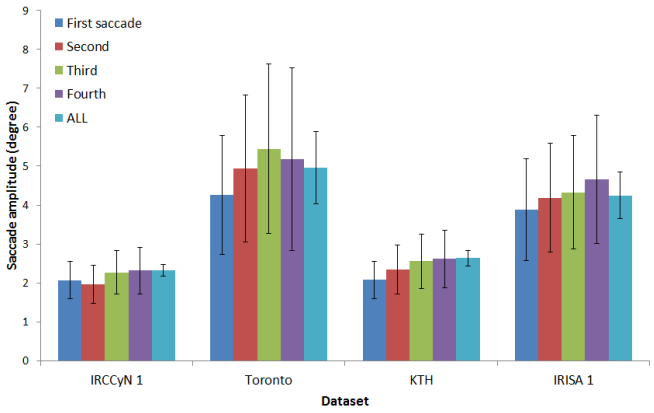
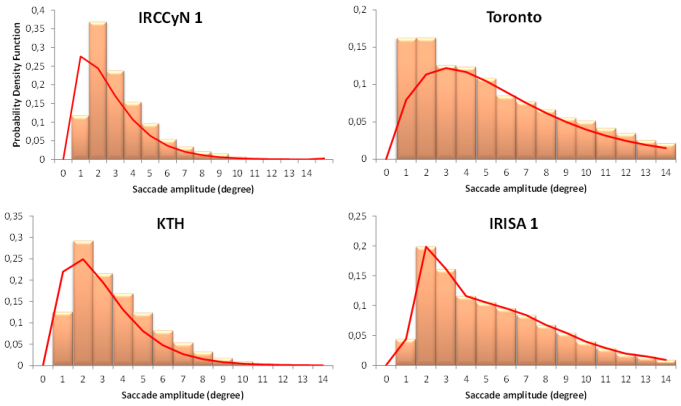


Table 2.2: Shape and scale of the Gamma distribution for IRCCyN's dataset in function of fixation rank (the label ALL means that all fixations have been taken into account). On the right-hand side, the Gamma distributions are plotted (see color correspondence).

leave one out) [165] and the method proposed by [23].
The first step of the one-against-all approach consists in computing a 2D fixation distribution from the fixation data of all observers except one for a given picture. The fixation distributions were then convolved with a two-dimensional Gaussian. Each pixel of this map represents the probability to be fixated. The standard deviation of the Gaussian kernel is set at one degree to reflect the foveal size. This map is then thresholded to select an image area having the highest probability of being fixated. The threshold is adaptively set in order to keep 25% of the image. The goal is now to compute the percentage of the visual fixations of the remaining observer that fall within salient parts of the thresholded saliency map. This process was iterated for all observers. For a given picture, the variability between observers is the average of the aforementioned percentage over all subjects. A value of 1 indicates that observers fixate the same areas. Conversely, a low value would suggest that the scan patterns are uncorrelated meaning a strong variability between subjects. This method is illustrated on figure 2.7.
A second method has been envisioned by Carmi and Itti [23].They assessed the fixation dispersion by defining the smallest bounding box enclosing all visual fixations. The ratio between the image surface and the bounding box provides a measure of fixation dispersion. This non parametric method has the advantage of simplicity. However, only one fixation occurring in the periphery of the images could have a strong impact on the dispersion estimation.
In the context of the evaluation of computational model of visual attention, the inter-observer congruency is a good indicator to gauge the complexity of the dataset. More precisely, if the inter-observer congruency is high, it indicates that there is something in the visual scenes that pops out, attracting observers' gaze. A computation model should be able to detect this area. In the other case, when there is nothing in the scene that stands out from the background, it becomes difficult or even impossible to predict accurately where observers look at in the scene.
Figure 2.8 (a) gives for four datasets the average congruency computed with the method one-against-all. Not surprisingly, the inter-observer congruency decreases over time. This is likely due to the fact that top-down influences in-



(a)



(b)

Figure 2.6: (a) Average saccade amplitudes for four state-of-the-art datasets. Error bars correspond to the confidence interval; (b) Saccade distributions for four datasets. The red curve is the Gamma distributions fitted to the raw data (without the quantization requested to build the histogram).

crease with the viewing time, as explained previously. Such influences increase the variability and the variety of visual strategies employed by observers. It is noticeable that the inter-observer congruency is different from one dataset to another. Figure 2.8 (b) illustrates the two top and bottom pictures having the highest and lowest inter-observer congruency. For a high congruency, there is something in the scene attracting our attention. For the first picture (top-left), there is a banner on the gate. Another top-down cue is present in the top-right picture. This is the horizon line which is a strong attractor of the visual attention [53]. Pictures for which the congruency is the lowest are illustrated on

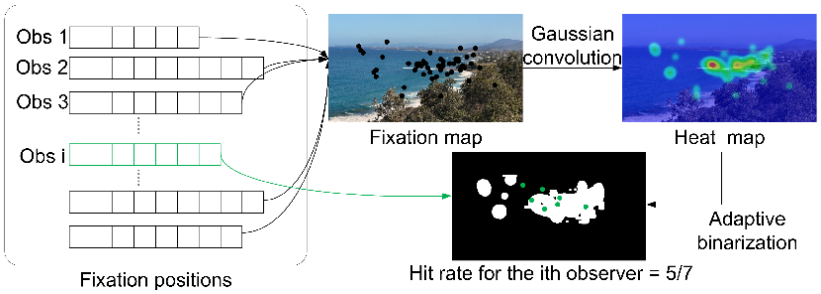
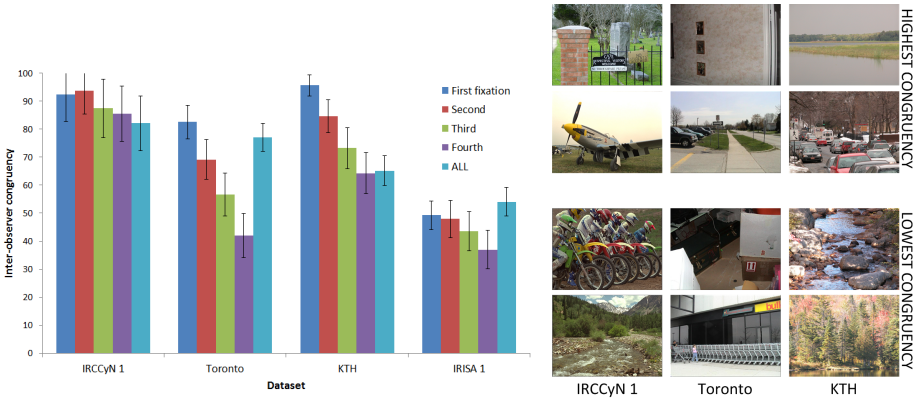


Figure 2.7: Inter-observer congruency computation.

the bottom of Figure 2.8 (b). On these pictures there are two reasons which could explain why the congruency is low. First is about the size or the number of salient regions as for the bikes or for the bottom-middle picture for which there are two distinct and distant salient regions (a pedestrian on the right and the yellow banner on the right). A high number of salient regions increases the fixation dispersion. The second reason is related to the "interestingness" of the scene. For the bottom-left and bottom-right pictures, there is nothing that really pops out, nothing able to steer our gaze in a particular direction. In this case, the congruency is also low. A score of inter-observer congruency is actually related to bottom-up and top-down cues, similarly to the visual salience. As far as we know, there is not study dealing with the evaluation of the contribution of bottom-up and top-down congruency. At this point, it is then difficult to conclude that the dataset having the lowest score is the most difficult one. As we explained before, the congruency depends on different low-level and high-level features.
Beyond this point, if we want to compare two datasets in terms of congruency, it is important to take into account the experimental protocols. It encompasses the number of pixel per degree (ppd) of visual angle, the number of observers, the viewing duration. For instance, the number of observers who participated to the elaboration of the dataset IRCCyN 1 [110] is twice the number of observers involved in the dataset Toronto [19]. Considering more observers statistically reduce the variability in the results.
For the sake of completeness, a third method should be mentioned. This is the visual clutter proposed by [149]. Contrary to the two other methods, the visual clutter relies only on the entropy of the low-level visual features and does not require the visual fixations.

Central bias

It is well known that observers' gaze is biased toward the screen's center (or stimuli). There are a number of causes of this central bias. Among them we just mention the photographer bias who tends to place the object of interest near the center's picture. A good review of central bias factor can be found



(a) (b)

Figure 2.8: (a) Inter-observer congruency over time for four eye tracking datasets. (b) the top first and bottom last pictures with the highest and the lowest congruency, respectively.

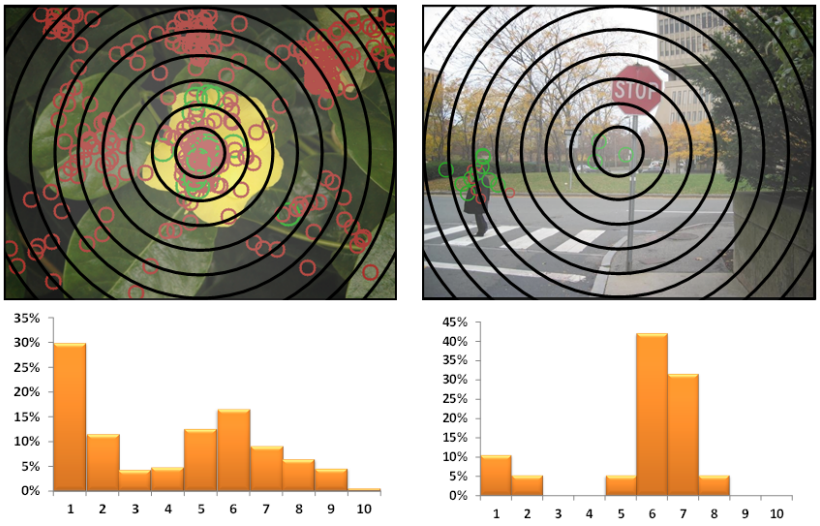


Figure 2.9: Illustration of the CBR method on two examples: right, a picture from the dataset KTH where the central bias is high; right, a picture from the dataset MIT CVCL. The central bias is low, observers had to look for pedestrian. On the bottom, the graph gives the distribution of visual points in function of the distance from the center.

in [167].
The central bias of a given dataset can be evaluated with two approaches: cor-

relation with a centered 2D Gaussian (called CCG) [110] and the center-bias ratio (called CBR) [14]. The former method consists in searching for the maximum linear correlation between a continuous saliency map and a centered 2D Gaussian. By varying the standard deviation of the Gaussian function, a more or less focused Gaussian is obtained. The standard deviation obtained for the maximum correlation score indicates the amount of center bias in the given image. For a high value of standard deviation, the center bias is small whereas a small value would indicate that there is a strong center bias. The second method [14] consists in generating a fixation map from all subjects. A set of 10 concentric circles is used (the radius is 10%, 20%,...,90% of the distance between the picture center and its top-left corner). The ratio of fixations falling within each crown (difference between two concentric successive circles) to the overall number of fixations is calculated. Figure 2.9 illustrates this approach. On the left-hand picture, the center bias is high (40% of the fixations are centered, enclosed by a circle having a radius of 20% of the distance between the center and the top-left corner) whereas the right-hand picture is not center-biased.
We evaluate the center-bias and how it evolves over time for 6 datasets. Figure 2.10 gives the standard deviation of the Gaussian function which maximizes the linear correlation coefficient (correlation is computed between the continuous saliency map and the 2D Gaussian function). As expected, results of the method CCG indicate that the center bias is more pronounced after the stimulus onset. Its influence decreases over time. There is again a high discrepancy between the datasets. DOVES and IRCCyN 1 present a high central bias. Toronto and KTH are less centred biased but still important. Indeed, the standard deviation is close to 6 degrees which is equivalent to 132 and 204 pixels around the center, respectively. At the opposite, the MIT CVCL dataset for which a pedestrian search task was given is not centre-biased, even for the first fixation. Results of the second method, CBR, confirm the previous conclusions. Figure 2.11 illustrates the CBR score over all visual fixations. Similar trends can be observed with slight differences. The most noticeable difference is about Toronto's dataset for which the center-bias is more pronounced (compared to those observed in figure 2.10).

2.3.3 Choosing appropriate test images

The evaluation and comparison of saliency models raise a number of issues such as the metric to use [104], the influence of high-level information and of course the set of images. This last point is fundamental since the dataset represents the ground truth from which the model performance is evaluated. We have mentioned in the first part of this chapter the most important factors (viewing duration, task, observers) for designing the experimental protocol. In the following, some recommendations are made about the images used for the eye tracking test.
First they should have the same resolution and the same onscreen size in order to subtend the same visual angle. This constraint is not mandatory but significantly eases the data processing. Indeed the use of the same resolution and

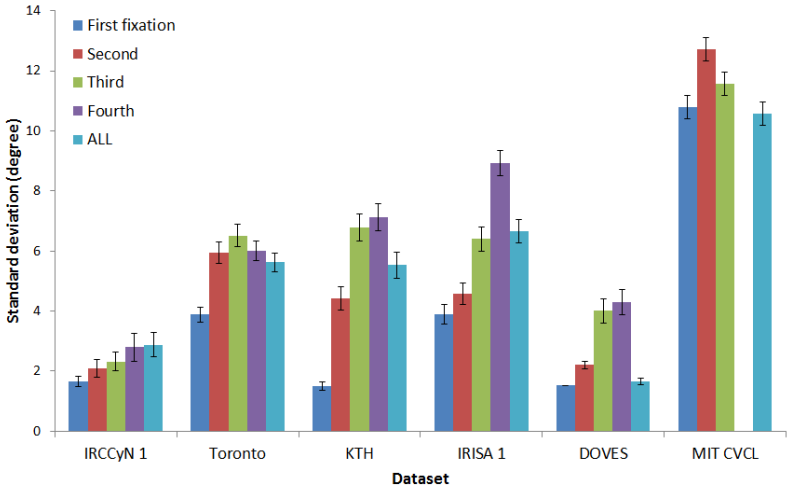


Figure 2.10: Standard deviation of the Gaussian function giving the highest linear correlation coefficient. The standard deviation is given in function of the fixation rank.

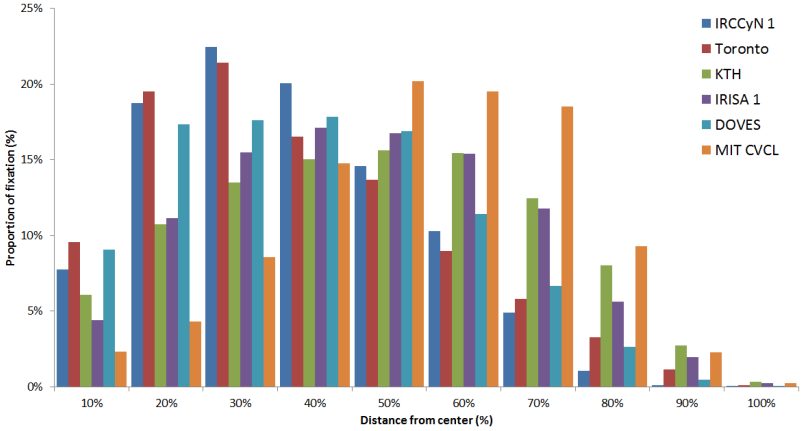


Figure 2.11: Center-bias analysis over 6 datasets. The percentage of fixations falling within a crown having a size of 10% of the distance between the top-left corner and the center is given.

display mode ensures a coherent nppd value for all pictures belonging to the dataset. If this is not the case, experimenter has to take some precautions:

- the threshold values used to extract the fixation from the raw eye tracking data should be adjusted per picture according to the visual angle;
- saccade amplitudes have to be normalized with the good nppd value;

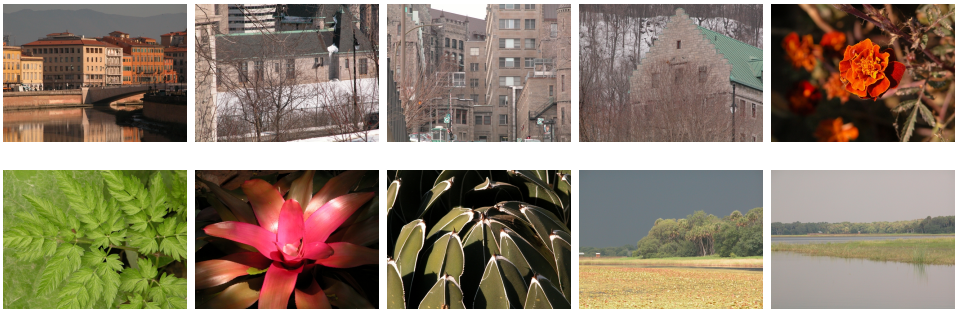


Figure 2.12: Ten pictures from the KTH dataset for which the inter-observer congruency is very low. In a context of evaluation of saliency models, these pictures could be considered as being useless.

- the computation of human saliency maps (obtained by equation 2.8) should take into account the appropriate standard deviation (σ) which reflects the fovea's size.

Second, the pictures should produce high inter-observer congruency. This means that the visual content must have a strong ability to attract our visual gaze on particular areas of the scene. Images for which the agreement between observers is low could be compared to noise; there is therefore no benefit to consider such content for the evaluation and comparison of attention models, given that the best prediction would be a random one. A method to compute the inter-observer congruency has been described and can be used to discard pictures having weak congruency. For instance, if we select the pictures having a congruency higher than 0.7, we would discard 20% and 69% pictures from Toronto and KTH dataset, respectively. Figure 2.12 illustrates ten pictures extracted from KTH datasets for which the agreement between observers is very low. As there is nothing salient in this kind of scenes, it is useless to predict where people would look at. Third, as mentioned in [14], the pictures with low-center bias should be preferred to those having a strong central bias. The fact that we tend to fixate more the screen's center than the periphery is a behavioral fact of our visual system [160]. However, some saliency models tend to favour by implementation the importance of the center of the picture, inducing a border effect. To deal with this issue, one solution would be to discard high-center bias pictures. In this paper, two methods have been described, one based on concentric circles (CBR) and the other based on the correlation between the saliency map and a centered 2D Gaussian (CCG). The former method is the simplest one but depends on the viewing condition (the radius of concentric circles is function of the picture's resolution). The latter method, explained in the previous section, is more appropriate since it relies on the nppd value.

2.4 Conclusion

Eye tracking dataset turns out to be a fundamental tool for vision research. This chapter provides some advices guiding researchers who want to create a new dataset for the evaluation and comparison of salient models. We list the main features of several existing datasets and examine some of them on the basis of different criteria. Two important points are underlined throughout this paper: the central bias and the dispersion between observers. We present and discuss methods to evaluate these two scene-based factors. A post-processing filtering could be used to discard pictures which present a strong central bias and/or a high dispersion between observers.
The software used to compute all information contained in this article is publicly available and can be re-used to reproduce tests. The software is available on the following link http://people.irisa.fr/Olivier.Le_Meur/publi/2012_BRM/index2.html.

Chapter 3

Similarity metrics

3.1 Introduction

Analysis of eye-tracking data has focused on synchronic indicators such as fixation (duration, number, etc) or saccade (amplitude, velocity, etc) rather than diachronic indicators (scanpaths or saliency maps). Synchronic means that an event occurs at a specific point in time, while diachronic means that this event is taken into account over time. We focus on diachronic measures, and review different ways of analysing sequences of fixations represented as scanpaths or saliency maps. Visual scanpaths depend on bottom-up and top-down factors such as the task users are asked to perform [157], the nature of the stimuli [186] and the intrinsic variability of subjects [174]. Being able to measure the difference (or similarity) between two visual behaviours is fundamental both for differentiating the impact of different factors and for understanding what govern our cognitive processes. It also plays a key role in assessing the performance of computational models on overt visual attention, by, for example, evaluating how well saliency-based models predict where observers look. In this chapter, we survey common methods for evaluating the difference/similarity between scanpaths and between saliency maps. We describe in Section 3.2 state-of-the-art methods commonly used to compare visual scanpaths. Section 3.3 presents the comparison methods which involve either two saliency maps or one saliency map plus a set of visual fixations. The strengths and weaknesses of each method are emphasized. The use of some of these metrics is illustrated in section 3.5. Finally some conclusions are drawn in section 3.6.

3.2 Methods for comparing scanpaths

Different metrics are available for comparing two scanpaths, using either distance-based methods (string edit technique or Mannan distance) or vector-based methods. Distance-based methods compare scanpaths only from their spatial characteristics, while vector-based approaches perform the comparison across different dimensions (frequency, time etc). These metrics are more or less complex and relevant depending on the situation to be analysed. However, there is no consensus in the community on the use of a given metric. In this section, we present three metrics: the string edit metric, Mannan's metric and a vector-based metric.

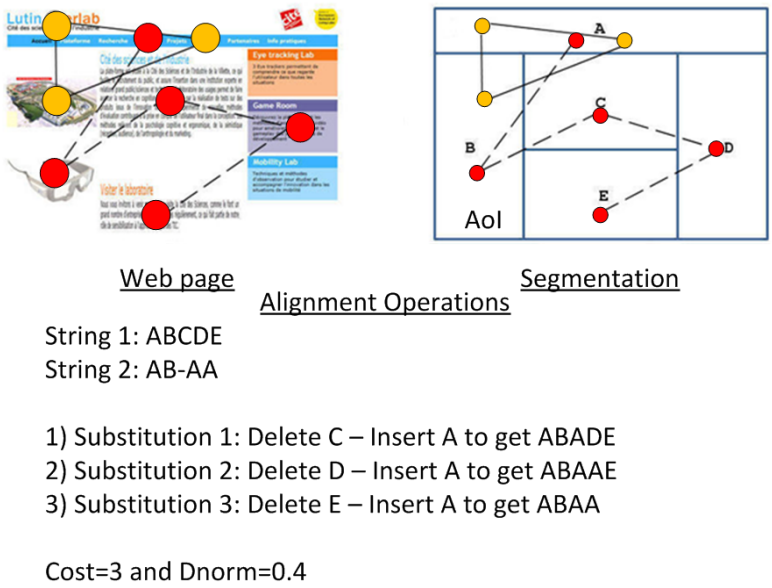


Figure 3.1: Computation of a string edit distance to align the sequences ABCDE and ABAA recorded on a web page. First, AOIs are segmented and coded by letters (A, B, C, etc). Second, the substitution operations are carried out. The total cost is equal to 3 (the minimum number of substitutions), and normalized to the length of the longer string, here 5, yielding an edit distance between the two strings of $d = (1 - \frac{3}{5}) = 0.4$.

3.2.1 String edit metric

The idea of the string edit metric is that a sequence of fixations on different Areas Of Interest (AOIs) can be translated into a sequence of symbols (numbers or letters) forming strings that are compared. This comparison is carried out by calculating a string edit distance (often called the Levenshtein distance) that gives a measure of the similarity of the strings [118]. This technique was originally developed to account for the edit distance between two words, and the measured distance is the number of deletions, insertions or substitutions that are necessary for the two words to be identical (which is also called the alignment procedure). This metric takes as input two strings (coding AOIs) and computes the minimum number of edits needed to transform one string into the other. A cost is associated with each transformation and each character. The goal is to find a serie of transformations that minimizes the cost of aligning one sequence with the other. The total cost is the edit distance between the two strings. When the cost is minimal, the similarity between the two strings is maximal (when the two strings are identical, the distance is equal to 0). Conversely, the distance increases with the cost and therefore with the dissimilarity between the two strings. Figure 3.1 illustrates the method. The Levenshtein distance is the most common way to compare scanpaths [87], [142] and has been widely used for assessing the usability of web pages [4].

The string edit distance can be computed using a dynamic programming technique (the WagFish algorithm [175]) that incrementally computes optimal alignments (minimizing the cost). The Levenshtein distance is not the only string edit distance that can be used for scanpaths. Others are described below:

- LCS is the length of the Longest Common Subsequence, which represents the score obtained by allowing only addition and deletion, not substitution.
- Damerau-Levenshtein distance allows addition, deletion, substitution and the transposition of two adjacent characters.
- Hamming distance only allows substitution (and hence, only applies to strings of the same length).

The advantage of the string edit technique is that it is easily computed and the order of fixations. It is also possible to compare observed scanpaths to predicted scanpaths when certain visual profiles are expected from the cognitive model used by the researcher. However, several drawbacks have to be underlined:

- Since the string edit is based on a comparison of a sequence of fixations occurring in pre-defined AOIs, the question is how to define these AOIs. There are two ways: automatically gridded AOIs or content-based AOIs. The former is built by putting a grid of equally sized areas across the visual material. For the latter the meaningful regions of the stimulus need to be subjectively chosen. Whatever AOIs are constructed, the string edit method means that only the quantized spatial position of the visual fixations are taken into account. Hence, some small differences in scanpaths may change the string while others produce the same string.
- The string edit method is limited when certain AOIs have not been fixated so there is a good deal of missing data.

3.2.2 Mannan's metrics

The Mannan distance [127],[128] is another metrics comparing scanpaths which are based on their spatial properties rather than their temporal dimensions, in the sense that the order of fixations is completely ignored. The Mannan distance compares the similarity between scanpaths by calculating the distance between each fixation in one scanpath and its nearest neighbour in the other scanpath. A similarity index I_s represents the average linear distance between two scanpaths D, with randomized scanpaths having the same size D_r. These randomly generated scanpaths are used for weighting the sequence of real fixations, taking into account the fact that real scanpaths may convey a randomized component. The similarity index I_s is given by

$$I_s = \left[1 - \frac{D}{D_r}\right] \times 100 \tag{3.1}$$

D is a measure of distance given by

$$D^2 = \frac{n_1 \sum_{j=1}^{n_2} d_{2j}^2}{2n_1 n_2 (a^2 + b^2)} + \frac{n_2 \sum_{i=1}^{n_1} d_{1i}^2}{2n_1 n_2 (a^2 + b^2)} \tag{3.2}$$

where,

- n_1 and n_2 are the number of fixations in the two traces.

- d_{1i} is the distance between the i^{th} fixation in the first trace and its nearest neighbor in the second trace.
- d_{2j} is the distance between the j^{th} fixation in the second trace and its nearest neighbor in the second one.
- a and b are the image's size.
- D_r is the distance between two sets of random locations.

The values returned by the algorithm I_s range from 0 (random scanpath) to 100 (identity). The main drawbacks of this technique are:

- The Mannan distance does not take into account the temporal order of fixation sequence. This means that two sequences of fixation having a reversed order but with an identical spatial configuration give the same Mannan distance.
- A difficult problem occurs when the two scanpaths have very different size (the number of fixations between them is very different). The Mannan distance may show a great similarity while the shapes of the scanpaths are definitely different. The Mannan distance is not tolerant to high variability between scanpaths.

3.2.3 Vector-based metrics

An interesting method was recently proposed by Jarodzka et al. [85]. Each scanpath is viewed as a sequence of geometric vectors that corresponds to subsequent saccades of the scanpath. The vector representation shows the length and the direction of each saccade. A saccade is defined by a starting position (fixation n) and ending position (fixation $n+1$). Then a scanpath with n fixations is represented by a set of $n-1$ vectors, and several properties can therefore be preserved, such as the shape of the scanpath, the scanpath length, and the position and duration of fixations. The sequences that have to be compared are aligned according to their shapes (although this alignment can be performed on other dimensions: length, durations, angle, etc). Each vector of one scanpath corresponds to one or more vectors of another scanpath, such that the path in the matrix of similarity between the vectors going from $(1,1)$ (similarity between the first vectors) to (n,m) (similarity between the last vectors) is the shortest one. Once the scanpaths are aligned, various measures of similarity between vectors (or sequences of vectors) can be used, such as average difference in amplitude, average distance between fixations and average difference in duration.

For example, figure 3.2 shows two scanpaths A and B (the first saccade is going upward). The alignment procedure attempts to match the five vectors (for the five consecutive saccades) of the participant scanpath with the four vectors of the model scanpath. Saccades 1 and 2 of scanpath A are aligned with saccade 1 of scanpath B, saccade 3A is aligned with saccade 2B, etc. Once the scanpaths are aligned, similarity measures are computed for each alignment. Jarodzka's procedure ends up with five measures of similarity (difference in shape, amplitude and direction between saccade vectors, distance between fixations and fixation durations).
This vector-based alignment procedure has a number of advantages over the string edit method. The first is that it does not need to determine pre-defined AOIs (and is therefore not dependent on the quantization of space). The second one is that it can align scanpaths not only on spatial dimension but also on any dimension available in saccade vectors (angle, duration, length, etc). For example, Lemaire et al. [115] used the spatial distance between saccades, the angle between saccades, and the difference of amplitude to realize the alignment. Thirdly, this alignment technique provides more

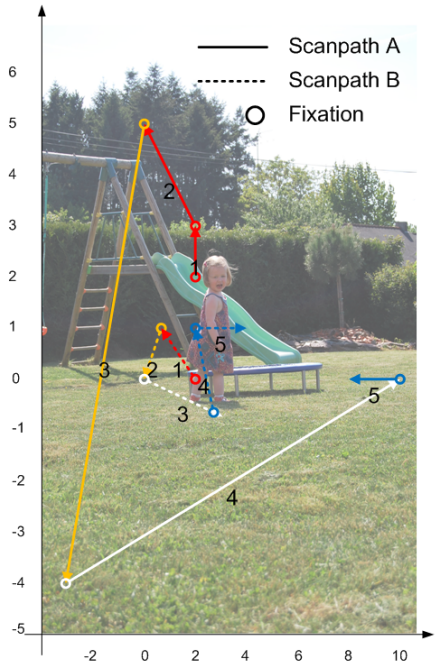


Figure 3.2: Alignment using saccadic vectors. The alignment procedure attempts to match the 5 vectors of the two scanpaths. The best match is the following: $1-2/1$; $3/2$; $4/3$; $5/4-5$.

detailed information on the type of (dis)similarity of two scanpaths according to the dimensions chosen. Lastly, this matrices deals with temporal issues, not only fixation durations, but it also successfully deals with shifts in time and variable scanpath lengths. The major drawbacks are:

- This measure only compares two scanpaths. Sometimes the overall aim is to compare whole groups of participants with each other.
- Eye movements such as smooth pursuit are not handled. Smooth pursuit movements are important when watching a video. However, the problem may be solved if it is possible to represent smooth pursuit as a series of short vectors which are not clustered into one long vector.
- The alignment procedure is independent of the stimulus content. However, the chosen dimensions may be weighted by some semantic values carefully selected by the researcher.

3.3 Methods for comparing saliency maps

Comparing two scanpaths requires to take a number of factors, such as the temporal dimension or the alignment procedure, into account. To overcome these problems,

another kind of method can be used. In this section, we focus on approaches involving two bi-dimensional maps. We describe three common methods used to evaluate the degree of similarity between two saliency maps: a correlation-based measure, the Kullback-Leibler divergence and ROC analysis.

3.3.1 Correlation-based measures

The Pearson correlation coefficient r between two maps H and P is defined as:

$$r_{H,P} = \frac{cov(H,P)}{\sigma_H \sigma_P} \tag{3.3}$$

where $cov(H,P)$ is the covariance between H and P, and σ_H and σ_P represent the standard deviation of maps H and P, respectively.
The linear correlation coefficient has a value between -1 and 1. A value of 0 indicates that there is no linear correlation between the two maps. Values close to zero indicate a poor correlation between the two sets. A value of 1 indicates a perfect correlation. The sign of r is helpful in determining whether data share the same structure. A value of -1 also indicates a perfect correlation, but the data vary together in opposite directions.
This indicator is very simple to compute and is invariant to linear transformation. Several studies have used this metric to assess the performance of computational models of visual attention [88], [110], [143]. Correlations are usually reported with degrees of freedom (the total population minus 2) in parentheses and the significance level. Note that the Spearman's rank correlation can also be used to measure the similarity between two sets of data [162].

3.3.2 The Kullback-Leibler divergence

The Kullback-Leibler divergence is used to estimate the overall dissimilarity between two probability density functions. Let us define two discrete distributions R and P with probability density functions r_k and p_k. The KL-divergence between R and P is given by the relative entropy of P with respect to R:

$$KL(R,P) = \sum_k p_k log \frac{r_k}{p_k} \tag{3.4}$$

The KL-divergence is only defined if r_k and p_k both sum to 1 and if $r_k > 0$ for any k such that $p_k > 0$.
The KL-divergence is not a distance, since it is not symmetric and does not satisfy the triangle inequality. The KL-divergence is non-linear. It varies in the range of zero to infinity. A zero value indicates that the two probability density functions are strictly equal. The fact that the KL-divergence does not have a well-defined upper bound is a strong drawback. In our context we have to compare two bi-dimensional saliency maps (H and P). We first transform these maps into two bi-dimensional probability density functions by dividing each location of the map by the sum of all pixel values. The probability that an observer focuses on position x is given by:

$$p_h(x) = \frac{H(x) + \epsilon}{\sum_i (H(i) + \epsilon)} \tag{3.5}$$

$$p_p(x) = \frac{P(x) + \epsilon}{\sum_i (P(i) + \epsilon)} \tag{3.6}$$

(a) Ground truth (b) 2% (c) 5% (d) 10% (e) 20%

Figure 3.3: Thresholded saliency maps to keep the top percentage of salient areas.

where ϵ is a small constant to avoid division by zero.

3.3.3 Receiver Operating Characteristic Analysis

The Receiver Operating Characteristic (ROC) analysis [62] is probably the most popular and most widely used method in the community for assessing the degree of similarity of two saliency maps. ROC analysis classically involves two sets of data: the first is from the ground truth (also called the actual values) and the second is the prediction (also called the outcomes).
Here we perform ROC analysis between two maps. It is also common to encounter a second method in the literature that involves fixation points and a saliency map. This method is described in section 4. Continuous saliency maps are processed as a binary classifier applied on every pixel. It means that the image pixels of the ground truth as well as those of the prediction are classified as fixated (or salient) or as not fixated (or not salient). A simple threshold operation is used for this purpose. However, two different processes are used depending on whether the ground truth or the prediction is considered:

- Thresholding the ground truth: the continuous saliency map is thresholded with a constant threshold in order to keep a given percentage of image pixels. For instance, we can keep the top 2, 5, 10, or 20% salient pixels of the map, as illustrated by Figure 3.3. This threshold is called $T^{x,G}$ (G for the ground truth and x indicating the percentage of image considered as being fixated).
- Thresholding the prediction: the threshold is systematically moved between the minimum and the maximum values of the map. A high threshold value corresponds to an over-detection whereas a smaller threshold affects the most salient areas of the map. This threshold is called $T^{x,P}$ (P for the prediction and x indicating the i^{th} threshold).

For each pair of thresholds, four numbers featuring the quality of the classification are computed. They represent the true positives (TP), the false positives (FP), the false negatives (FN) and the true negatives (TN). The true positive number is the number of fixated pixels in the ground truth that are also labelled as fixated in the prediction.

Figure 3.3 illustrates the thresholding operation on the Parrot picture. The first continuous saliency map (b) of Figure 3.3 is thresholded to keep 20% of the image ($T^{20,G}$) and is compared to the second continuous saliency map (b) of Figure 3.4. The classification result is illustrated by Figure 3.4. The red and uncoloured areas represent pixels having the same label, i.e. a good classification (True Positive). The green areas

(a) 20% of ground truth (b) Prediction (c) Classification

Figure 3.4: Classification result (on the right) when considering a 20% thresholded ground truth (left picture) and a prediction (middle picture). Red areas are True Positives, green areas are False Negatives, and blue areas are False Positives. Other areas are True Negatives.

represent the pixels that are fixated but are labelled as non-fixated locations (False Negative). The blue areas represent the pixels that are non-fixated but are labelled as fixated locations (False Positive). A confusion matrix is often used to visualize the algorithm's performance (see Figure 3.5(c)). An ROC curve that plots the false positive rate as a function of the true positive rate is usually used to display the classification result for the set of thresholds used. The true positive rate (TPR), also called sensitivity or recall, is defined as $TPR = TP/(TP + FN)$, whereas the false positive rate (FPR) is given by $FPR = FP/(TP + FN)$. The ROC area or the AUC (Area Under Curve) provides a measure indicating the overall performance of the classification. A value of 1 indicates a perfect classification. The chance level is 0.5. There are different methods to compute the AUC. The simplest ones are based on the left and right Riemann sums. The left Riemann sum is illustrated by Figure 3.5. A more efficient approximation can be obtained by a trapezoid approximation: rather than computing the area of rectangles, the AUC is given by summing the area of trapezoids. In our example, the AUC value is 0.83.

3.4 Hybrid method

So far we have focused on similarity metrics involving two scanpaths or two saliency maps. In this section we describe methods based on a saliency map and a set of fixation points. We call this kind of method hybrid as it mixes two types of information. Four approaches are presented: ROC analysis, Normalized Scanpath Saliency, percentile and the Kullback-Leibler divergence.

3.4.1 Receiver Operating Characteristic Analysis

The ROC analysis is performed here between a continuous saliency map and a set of fixations. The method tests how the saliency at the points of human fixation compares to the saliency at non-fixated points. As in the previous section, the continuous saliency map is thresholded to keep a given percentage of pixels of the map. Each pixel is then labelled as either fixated or not-fixated. For each threshold the observers fixations are laid down on the thresholded map. The true positive (fixations that fall on fixated areas) and the false negative (fixations that fall on non-fixated areas) are determined (as illustrated by Figure 3.6). A curve that shows the TPR (or hit rate)

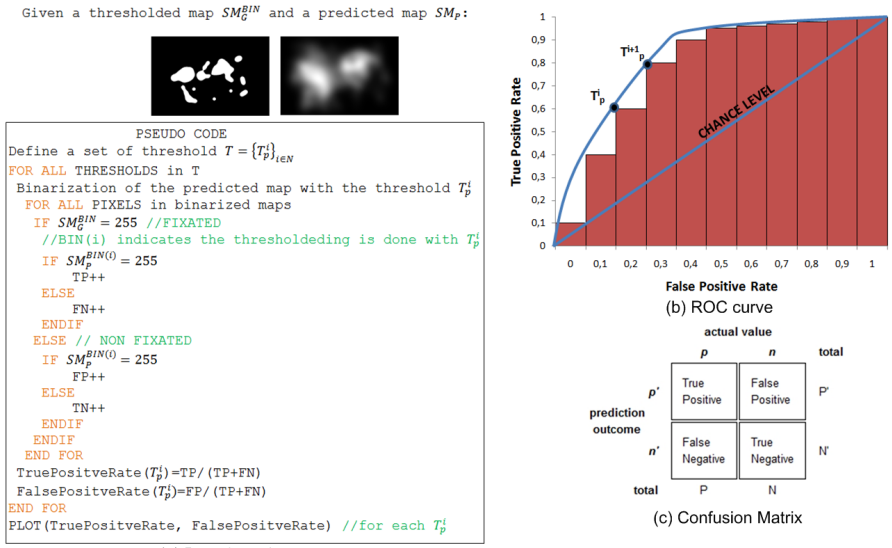


Figure 3.5: Pseudo code to perform an ROC analysis between two maps (a), ROC curve (b) and the confusion matrix (c). The AUC is approximated here by a left Riemann sum as illustrated in (b).

(a) Perfect Hit (b) 50% of hit

Figure 3.6: Example of ROC analysis. Red and green dots are the fixations of two observers for the Parrots image. These dots are drawn on a thresholded saliency map. On the left hand side the hit rate is 100% whereas the rate is 50% for the example on the right hand side.

as a function of the threshold can be plotted. Note that the percentage of the image considered to be salient is in the range of 0 to 100%. If the fixated and non-fixated locations cannot be discriminated, the AUC will be 0.5. This first analysis method is used in papers such as [161] and [165]. Although interesting, this method is not sensitive to the false alarm rate.

To deal with the previous limitation, a set of control points, corresponding to

non-fixated points, can be generated. Two methods commonly encountered in the literature are discussed. The first method is the simplest one and consists in selecting control points from either a uniform or a random distribution. This solution does not take into account the fact that fixations are distributed neither evenly nor randomly throughout the scene. The second method, proposed by [43], [161] defines control points by choosing locations randomly from a distribution of all fixation locations that occurred at the same time, but on other images. This way to define the control point is important for different reasons. First, as the fixations come from the same observer, so the same bias, systematic tendency or peculiarity of the observer are taken into account. These factors then have a limited influence on the classification results. Among them, the most important influence is the central bias. A number of factors can explain this central tendency (see Chapter 2, section 2.3.2) Secondly, the set of control points has to stem from the same time interval as the set that is analysed. Indeed, bottom-up and top-down influences are not similar over time. For instance, bottom-up influences are maximal just after the stimulus onset. Top-down influences tend to increase with viewing time, leading to a stronger dispersion between observers. Although the second method is more robust than the first one, the method has a serious flaw. It underestimates the salience of areas which are more or less centred in the image.
In a similar fashion to the method in section 3.4, the control points and the fixation points are then used to plot a ROC curve. For each threshold the observer's fixations and the control ones are laid down on the thresholded map. The true positive rate (fixations that fall on fixated areas) and the false positive rate are determined. From this ROC curve the AUC is computed. The confidence interval is computed by using a non-parametric bootstrap technique [40]. Many samples having the same size as the original set of human fixations are generated by sampling with replacement. These samples are called bootstrap samples. In general 1,000 bootstrap samples are created. Each bootstrap sample is used as a set of control fixations. The ROC area between the continuous saliency map and the points of human fixation plus the control points is computed. The bootstrap distribution of each ROC analysis is computed and a bootstrap percentile confidence interval is determined by percentiles of the bootstrap distribution, leaving off $\alpha/2 \times 100$ of each tail of the distribution where α is the confidence level.
Sometimes, the quality of the classification relies on the equal error rate (EER). The equal error rate is the location on an ROC curve where the false positive rate and the true positive rate are equal (i.e. the error at which false alarms equal the miss rate $FPR = 1 - TPR$). As with the AUC, the EER is used to compare the accuracy of the prediction. In general, the system with the lowest EER is the most accurate.

3.4.2 Normalized scanpath saliency

The Normalized Scanpath Saliency [140] is a metric involving a saliency map and a set of fixations. The idea is to measure the saliency values at fixation locations along a subject's scanpath.
The first thing to do is to standardize the saliency values in order to have a zero mean and unit standard deviation. It is simply given by

$$Z_{SM}(\mathbf{x}) = \frac{SM(\mathbf{x}) - \mu}{\sigma} \tag{3.7}$$

where Z_{SM} is the standardized saliency map and

$$\mu = \frac{1}{|I|}\sum_{\mathbf{x}\in I} SM(\mathbf{x}) \tag{3.8}$$

$$\sigma = \sqrt{\frac{1}{|I|}\sum_{\mathbf{x}\in I}(SM(\mathbf{x})-\mu)^2} \tag{3.9}$$

where $|I|$ indicates the number of pixels of the picture I. For a given coordinate, the quantity $Z_{SM}(\mathbf{x})$ represents the distance between the saliency value at $\mathbf{x}$ and the average of saliency expressed in units of the standard deviation. This value is negative when the saliency value at the fixation locations is below the mean, positive when above. To take account of the fact that we do not focus accurately on a particular point, the NSS value for a given fixation location is computed on a small neighbourhood centred on that location:

$$NSS(x_{f^{(k)}}) = \sum_{\mathbf{x}\in\pi} K_h(x_{f^{(k)}} - \mathbf{x}) Z_{SM}(\mathbf{x}) \tag{3.10}$$

where K is a kernel with a bandwidth h and π is a neighbourhood.
The NSS is the average of $NSS(x_{f^{(k)}})$ for all fixations M of an observer. It is given by

$$NSS = \frac{1}{M}\sum_{k=1}^{M} NSS(x_{f^{(k)}}). \tag{3.11}$$

Figure 3.7 illustrates the computation of the NSS value for a scanpath composed of 8 visual fixations. In this example, the average NSS value is 0.3, indicating a good correspondence between the model-predicted saliency map and the observer's scanpath.

3.4.3 Percentile

In 2008, Peters and Itti designed a metric called percentile [139]. A percentile value $P(x_{f^{(k)}})$ is computed for each location of fixation points $x_{f^{(k)}}$. This score is the ratio between the number of locations in the saliency map with values smaller than the saliency value at point $x_{f^{(k)}}$ and the set of all locations. The percentile value is defined as follows:

$$P(x_{f^{(k)}}) = 100 \times \frac{|\{\mathbf{x}\in X : SM(x) < SM(x_{f^{(k)}})\}|}{|SM|} \tag{3.12}$$

where X is the set of locations of the saliency map SM and $x_{f^{(k)}}$ is the location of the k^{th} fixation. $|.|$ indicates set size.
The final score is the average of $P(x_{f^{(k)}})$ for all fixations of an observer. By definition, the percentile metric has a well-defined upper bound (100%) indicating the highest similarity between fixation points and saliency map. The chance level is 50%.

3.4.4 The Kullback-Leibler divergence

The KL-divergence, defined in section 3.3.2, is used here to compute the dissimilarity between the histogram of saliency sampled at eye fixations and that sampled at random locations. Itti and Baldi [82] were the first to use this method. The set of control

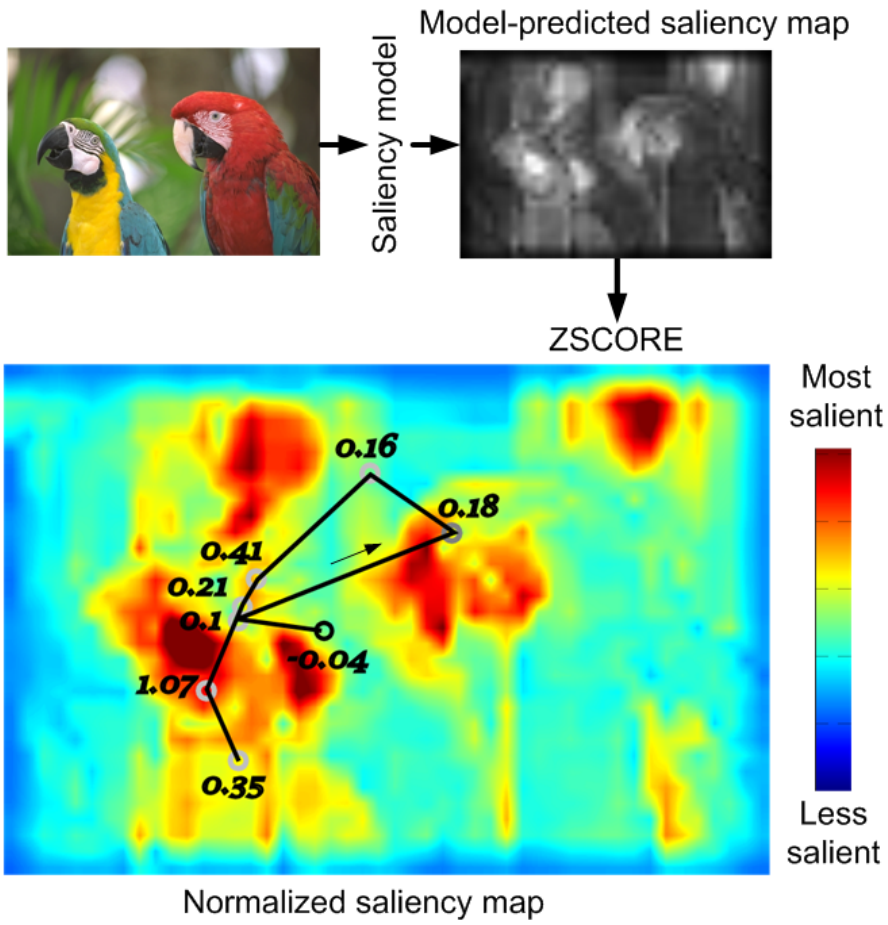


Figure 3.7: Example of NSS computation: the heat map is a normalized version of the model-predicted saliency map with a zero mean and unit standard deviation. A scanpath composed of 8 fixations (grey circles; the black one is the first fixation) is overlaid upon the standardized map. The normalized salience is extracted for each location. Values are shown in black next to the fixations.

points (or the set of non-fixated points) are drawn from a uniform spatial distribution. However, human fixations are not randomly distributed, since they are governed by various factors such as the central bias explained earlier. To be more agnostic to this kind of mechanism, Zhang et al. [189] measured the KL-divergence between the saliency distribution of fixated points of a test image and the saliency distribution at the same pixel locations but of a randomly chosen image from the test set. To evaluate the variability of the score, the evaluation was repeated 100 times with 100 different sets of control points.
Contrary to the previous KL-divergence method of section 3.3.2, a good prediction has

Model - Dataset	IRCCyN 1	Toronto
Itti	0.60±0.10	0.99±0.05
Le Meur	0.77±0.13	0.87±0.03
Bruce	0.60±0.09	0.72±0.04
Judd	0.82±0.11	0.87±0.05

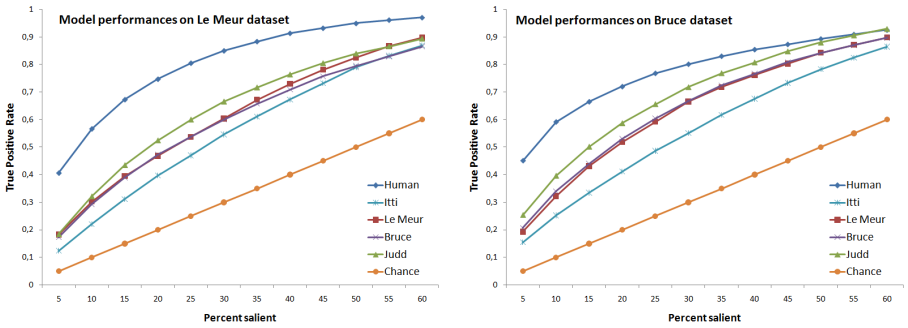
Table 3.1: NSS scores for four state-of-the-art saliency models on the Le Meur and Bruce datasets ($AVG \pm SEM$). SEM is the Standard Error of the Mean. A high average NSS value indicates a good prediction. Itti's model performs much better on Bruce's dataset than on Le Meur's dataset. Judd's model gives similar results for both datasets.

a high KL-divergence score. Indeed, as the reference distribution represents chance, the saliency computed at human-fixated locations should be higher than that computed at random locations.

3.5 Benchmarking computational models

Performance of the most prominent saliency models is here examined. The quality of the predicted saliency maps is given here by two metrics: the hit rate and the NSS. These metrics are hybrid metrics, since they involve a set of visual fixations and a map. We believe that these metrics are the best way to assess the relevance of a predicted saliency map. Compared to saliency map-based methods, hybrid methods are non-parametric. Human saliency maps are obtained by convolving a fixation map by a 2D Gaussian function, which is parametrized by its mean and its standard deviation. Note that instead of using the hit rate, we could have used an ROC analysis. To perform the analysis, we use two eye-tracking datasets that are available on the Internet (called IRCCyN1 and TORONTO, see Table 2.1 for more details).
We compare the performance of four state-of-the-art models: Itti's model [84], Le Meur's model [110], Bruce's model [19] and Judd's model [92].
Figure 3.8 gives the ROC curve indicating the performance of different saliency models averaged over all testing images. The method used here is the method described at the beginning of section 3.4.1. The upper-bound, i.e. the inter-observer variability, was computed by the method proposed by [165] and described in section 2.3.2. Table 3.1 gives the average NSS value over the two tested datasets.
Under the ROC metric, Judd's model has the highest performance, as illustrated by Figure 3.8. This result was expected, since this model uses specific detectors (face detection for instance) that improve the ability to detect salient areas. In addition, this model uses a function to favour the centre of the picture in order to take the central bias into account. However, the results are more contrasted under the NSS metric shown in Table 3.1. On average across both databases, Judd's model is still the highest performing. On Bruce's dataset, Itti's model performs the best, with a value of 0.99, whereas Judd's model performs at 0.87. The model ranking is therefore dependent on the metric used. It is therefore fundamental to use more than one metric when assessing the performance of computational models of visual attention.

Remark: a more extensive benchmark has been independently performed and released in 2012 [14]. The model we proposed in 2006 is ranked according to three metrics



(a) (b)

Figure 3.8: Models performance tested on (a) IRCCyN1 dataset; (b) TORONTO dataset. All models perform better than chance and worse than humans. Judd's model gives the best performance on average.

Metric - Dataset	TORONTO	KTH	MIT CSAIL
CC	11	7	8
NSS	10	9	9
AUC	10	15	17

Table 3.2: Ranking of our visual attention model over three datasets (TORONTO, KTH, MIT CSAIL) and 28 visual attention models. Deduced from figure 7 of [14].

(CC, NSS and AUC). Twenty-height models have been considered. Our model is in the top third of the ranking, as given by table 3.2. This is a good score considering the fact that the model has been designed in 2005-2006.

3.6 Conclusion

This chapter provides an extensive overview of the different ways of analysing diachronic variables from eye-tracking data, because they are generally under-used by researchers. These diachronic indicators are scanpaths or saliency maps generated to represent the sequence of fixations over time. They are usually provided by eye-tracking software for illustrative purposes, but no real means to compare them are given. This chapter aims to fill that gap by providing different methods of comparing diachronic variables and calculating relevant indices that might be used in experimental and applied environments. These diachronic variables give a more complete description of the visual attention time course than synchronic variables, and may inform us about the underlying cognitive processes. The ultimate step would be to relate the visual behaviour recorded with eye-trackers accurately to the concurrent thoughts of the user.

3.7 Contribution in this field

The main contribution in this field consists of a survey of methods used to assess the degree of similarity between eye tracking data and a prediction. Our contribution is an article published in the journal Behavior Research Methods [Impact Factor=1.907]. It covers different aspects of cognitive science and includes a particular focus on use of computer technology in psychological research. Along this publication, we release a software which implements most of the methods described in the article. At this moment, the software has been downloaded more than fifty times.
Journal:

- O. Le Meur and T. Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behavior Research Method, 2012.

Software:

- VisualFixationAnalysis software: user-interface released to the community.

Part II

Attention-based applications

Chapter 4

Quality assessment

4.1 Introduction

A great deal of interest and research has been devoted to the design and development of visual quality metrics, leading to the definition of three types of quality metrics: no-reference, reduced-reference and full-reference video quality metrics. A full-reference video quality metrics requires to have the original and the impaired video sequences. This is obviously a strong limitation in practice. To overcome this limitation, a reduced-reference quality metrics can be used. It requires to get a reduced description of the reference video. This reduced description is compared to a similar description extracted from the impaired video to infer a quality score. The more the descriptions are close, the higher the quality is. For some application such as monitoring the quality in a transmission chain, this kind of approach is much more convenient than a full-reference video quality. However, one drawback is that the description extracted from the original video sequence must be encoded without loss. The last solution is to use a no-reference quality metrics for which only the impaired video sequence is available. No-reference quality metrics are less complex but less powerful.
The most relevant quality metrics (IQM (Image Quality Metric) or VQM (Video Quality Metric)) use human visual system properties to predict accurately the quality score that an observer would have given. Hierarchical perceptual decomposition, contrast sensitivity functions, visual masking, etc are the common components of a perceptual metric. These operations simulate different levels of human perception and are now well mastered. In this chapter, we present quality metrics using visual attention. Section 4.2 presents the link that might exist between visual attention and quality score. In Sections 4.3 and 4.4, we examine the influence of video coding artefacts as well as the task of subjective quality assessment on the visual deployment, respectively. Section 4.5 describes saliency-based pooling methods used to combine distortion values into an unique quality score. Finally we conclude in Section 4.6.

4.2 Visual attention and quality assessment

Assessing the quality of an image or video sequence is a complex process, involving the visual perception as well as the visual attention. It is wrong to think that all areas of

the picture or video sequence are accurately inspected during a quality assessment task. People preferentially and unconsciously focus on regions of interest. For these types of regions, our sensitivity to distortions might be significantly increased compared to non-salient regions. Even though we are aware of this, very few IQM or VQM approaches take this property into account. To go one step further on this topic, we performed several experiments in order to understand how people perceive the quality of a video and how they adapt their visual strategies to judge the quality of an image or video sequence. For continuous quality evaluation, we know that humans are quick to criticize and slow to forgive. This experimental property can be used to improve the pooling stage of video quality metrics. However, there is almost no study related to the visual strategy of an observer during a quality assessment task. It is intuitively obvious that the areas of the video sequence do not have the same visual importance and the same capacity to draw our visual attention. The hypothesis is that an impairment appearing on a region of interest is probably more annoying than an impairment on a non visually interesting area. Is this intuition relevant and does the use of the visual importance of an area bring a significant improvement? Previous studies dealing with the quality assessment of still color pictures [130] showed that the relationship between visual importance and the quality assessment is not as simple as one would expect.

4.3 Do video coding artifacts influence our visual attention?

In [111], we investigated whether the presence of strong visual coding degradations disturbed the deployment of visual attention in a free-viewing task. To reach this objective, eye-tracking experiments were performed on video sequences with and without video coding artefacts. Observers were asked to watch the video clips without specific instruction.

We found out that the saliency sequences for the impaired sequences are not significantly different from the original ones. Although that the degradations of the video clips were at least estimated as annoying by a panel of observers, the visual attention is almost invariant to video coding artefacts (impairments affect attention but the effect is rather small). Considering that the deployment of the visual attention is significantly influenced by the low-level visual properties (especially under free viewing) and that the quality of the video was significantly reduced (to be at least annoying when a specific task of quality was given), it was not absurd to presume that observers would watch the video clips in a different way than those watching the same unimpaired clips. This is not the case, even though great care was taken on the way the quality of the video sequences was degraded. Indeed the amount of impairment was not at all uniformly distributed spatially as well as temporally. It was expected that these variations of quality disturb the attention of the observer. How could we explain that there is only few modifications of the overt visual attention?

This result would indicate that the oculomotor behaviour is also influenced by factors others than the low-level visual features, under free viewing task. It is not surprising since the transformation of visual precepts is the result of a series of complex biological and mental processes. As stated by Lester [117], visual perception is a function of the meaning we associate -through learned behaviour or intelligent

assumptions- with the object we see.
In addition the fact that there is no explicit task does not mean that top-down influences are ruled out. To catch a total comprehension and understanding of visual images, observers use their own knowledge (memory, shape recognition...) to understand, to recognize and to interpret the scene. No one can dispute the importance of early vision (see section 2.2.3 for more details). Human fixations (or saliency maps) can be predicted by mathematical models. As described in Chapter 1, several models, purely based on the low-level visual features, exist in the literature. They perform quite well but could be greatly improved when a combination of contextual information and low-level visual features is used [165]. This simply demonstrates that the source guidance of our visual attention is related to low-level visual features but not exclusively.
Finally, the fact that there is no significant modification in the deployment of visual attention in presence of distortion would suggest again that the fixation points are closely linked with the semantic and the context of the scene semantic, as suggested by [70]. However, it could be argued object's shape are not sufficiently degraded both to impair the shape/pattern recognition and to disturb the scene understanding. Conversely to fidelity metric, Rouse and Hemami [152] introduced the concept of similarity metric. This kind of metrics assesses the quality of edges of the shapes in order to score the visual equivalence between two images. This score also indicates the usefulness or utility of the content. An extension of our study could consist in impairing the set of video sequences to dramatically reduce their utility scores. A new eye tracking experiment would be required to evaluate the visual influence of coding artefacts.

4.4 Free-viewing vs quality-task?

In [112], we investigated the influence of quality assessment task on the visual deployment. To understand how people watch a video sequence during quality assessment and free-viewing tasks, two eye tracking experiments were carried out.

The comparison between gaze allocations indicates the quality task has a moderate impact on the visual attention deployment. A first test performed on the fixation durations does not reveal a significant difference between the two conditions (free-task vs quality-task). A second test consisted in comparing the human saliency maps. The degree of similarity between these maps were evaluated by using a ROC analysis and by computing the area under curve (AUC) (see section 3.3.3 for ore details). The AUC computed for each frame were then averaged over the video sequence to get the final similarity score. The similarity degree between the human priority maps (free-task vs quality-task) is high (greater than 0.85 in average). This would indicate eye movements are significantly influenced by neither the level of impairment nor the quality task. However, when the number of presentation increases, the similarity decreases indicating the presence of a memory or learning effect.

4.5 Saliency-based quality metric

The two experiments described in previous sections suggest the visual deployment for a free-viewing and quality tasks is not significantly different. This indicates that the saliency areas can be predicted in both conditions by a purely bottom-up computa-

tional model of visual attention. From the predicted saliency map, we can now check the assumption that degradations on salient areas are more annoying than other distortions.

We adapted an objective full-reference video quality metric we previously designed (see [131] for details of the WQA metrics (Wavelet-based QuAlity metrics)). The modification consists in taking into account the visual importance of the video sequence areas in the pooling function:

$$D_t = \left(\frac{\sum_{k=1}^{K} \sum_{l=1}^{L} w_i(x,y,t) \cdot \left(d(x,y,t) \right)^{\beta_s}}{\sum_{k=1}^{K} \sum_{l=1}^{L} w_i(x,y,t)} \right)^{\frac{1}{\beta_s}}, \tag{4.1}$$

Where D_t is the perceptual distortion value for the frame at time t weighted by the visual saliency. K and L are the height and the width of the image, respectively. $w_i(x,y,t)$ is the weighting factor i applied at pixel (x,y) of the frame at time t. $d(x,y,t)$ is the spatio-temporal map of the visual distortion at t. For more details readers could refer to [131]. Two β_s values were tested: 1 and 2. A higher β_s value will favour the strongest distortions into the frame to the detriment of others. Seven different weighting functions w_i have been defined and tested:

$$\left| \begin{array}{l} w_0(x,y,t) = 1 \\ w_1(x,y,t) = SM_n(x,y,t) \\ w_2(x,y,t) = 1 + SM_n(x,y,t) \\ w_3(x,y,t) = SM(x,y,t) \\ w_4(x,y,t) = 1 + SM(x,y,t) \\ w_5(x,y,t) = SM_b(x,y,t) \\ w_6(x,y,t) = 1 + SM_b(x,y,t) \end{array} \right. \tag{4.2}$$

where $SM(x,y,t)$ is the unnormalized human saliency map, $SM_n(x,y,t)$ is the human saliency map normalized in the range $[0,1]$ and $SM_b(x,y,t)$ is a binarized human saliency map. We remind the saliency maps are computed from eye data collected during the experiment involving the impaired video sequence in quality task. The weighting functions w_1, w_3 and w_5 give more importance to the salient areas than the others. Indeed, the offset value of 1 in the weighting functions w_2, w_4 and w_6 allows us to take into account distortions appearing also on the non salient areas. w_0 is the baseline quality metrics in which the pooling is not modified. The final distortion value D, pooled over the sequence, is obtained by the formula (7) given in [131].

The impact of each weighting function was evaluated using the linear correlation coefficient (CC), the Spearman rank ordered correlation coefficient (SROCC) and the Root Means Squared Error (RMSE) between the MOS (Mean Opinion Score) and its prediction $MOSp$ (Predicted MOS). All the results are given in Table 4.1. Whatever the weighting functions used, there is no significant performance improvement. The best results are obtained with a constant weighting w_0, meaning that all the areas of the video sequences are considered as having the same visual importance. These results suggest that a simple saliency-based pooling function is not a good solution to improve the visual quality prediction. Our initial assumption is here not verified.

Weighting			Metrics		
Saliency	w_i	β_s	CC	SROCC	RMSE
IMP(QT)	w_0	***1***	***0.889***	***0.904***	***0.526***
	w_1	1	0.875	0.903	0.554
	w_2	1	0.889	0.904	0.525
	w_3	1	0.875	0.903	0.554
	w_4	1	0.883	0.908	0.538
	w_5	1	0.876	0.904	0.553
	w_6	1	0.89	0.906	0.524
IMP(QT)	w_0	***2***	***0.892***	***0.9***	***0.519***
	w_1	2	0.878	0.904	0.548
	w_2	2	0.892	0.901	0.519
	w_3	2	0.878	0.904	0.548
	w_4	2	0.886	0.912	0.532
	w_5	2	0.88	0.905	0.546
	w_6	2	0.893	0.902	0.517

Table 4.1: Impact of the human saliency on the performances of a video quality metric. Different weighting functions are used.

4.6 Conclusion

In this chapter, we had studied and analysed the visual deployment in a context of quality assessment. The first experiment evaluates whether or not visual coding artefacts modify the way we look within a scene. The second experiment aims at comparing the visual deployment between a free-viewing and a quality assessment task. These two experiments were designed to support the idea of using a purely bottom-up model of visual attention in order to weight the visual distortion in function of their visual importance.
The comparison between eye movements collected during these experiments indicates that the degree of similarity between human priority maps (original vs impaired; free-viewing vs quality-task) is very high. Two conclusions can be drawn from these experiments:

- the gaze allocation is not disturbed by the level of distortion in a free-viewing task.
- when observers were instructed to score the quality of the video sequences, their gaze deployments were not significantly different from the gaze deployments observed without task (free-viewing).

The previous conclusions give support to the use of a bottom-up computational model of visual attention to steer the pooling method of quality metrics.

The video quality metric of [131] has been modified in order to take into account the visual importance of the areas of the impaired video sequence. Different weighting functions based on the human saliency maps have been proposed. Neither of them succeeds in improving the performance of the quality metric. The hypothesis postulating that impairments contribute more to the elaboration of the quality score when

it occurs on a region of interest is likely true, but the proposed strategy is probably not appropriate. What is certain is that observers have to inspect some areas more or less accurately in order to assess the video quality. Among all the visual fixations, some of these fixations contribute to the quality assessment whereas others have low or no impact. A new avenue of investigation would be to examine the relationship between the duration of the visual fixations and the amount of distortion. The idea is that observers do not require to focus a long time on strong distortions whereas, when the amount of the distortion of an area is small, observers need more time to inspect and to judge the quality. This new hypothesis could be of strong importance since the definition of the saliency would be drastically modified.

4.7 Contributions in this field

Journal:

- O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, Do video coding impairments disturb the visual attention deployment?, Elsevier, Signal Processing: Image Communication, vol. 25, Issue 8, pp. 597-609, September 2010.
- O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, Overt visual attention for free-viewing and quality assessment tasks. Impact of the regions of interest on a video quality metric, Elsevier, Signal Processing: Image Communication, vol. 25, Issue 7, pp. 547-558, August 2010.
- A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, Considering temporal variations of spatial visual distortions in video quality assessment, IEEE Journal of Selected Topics in Signal Processing, Special Issue On Visual Media Quality Assessment, vol. 3, Issue 2, pp. 253-265, 2009.

Conferences:

- A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Which Semi-Local Visual Masking Model For Wavelet Based Image Quality Metric?, ICIP, 2008.
- A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric, ICIP, 2007.
- A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Task impact on the visual attention in subjective image quality assessment, EUSIPCO, 2006.

Chapter 5

Memorability

5.1 Introduction

The study of images memorability in computer science is a recent topic [79, 95]. From those first attempts it appears that it is possible to predict the degree of image's memorability quite well. Learning algorithms have been used to infer from a set of low-level visual features the extent to which an image is memorable. Although Isola et al. [79] expressed the intuition that memorability and visual attention might be linked, they did not study further this relationship. Khosla et al. [95] proposed a local descriptor based on Itti's model [84]. The performance of this descriptor alone is low. In this chapter we intend to show that attention-based cues and features might have high importance in memorability both from a experimental and predictive point of views. In the next sections we will focus on an eye-tracking experiment using images from Isola's database and the cues which can be extracted from gaze behaviour and which might be related to the memorability score of the images. In section 5.3, we evaluate the relevance of two attention-related features and show that by using the same classifier we obtain comparable and even better memorability results than [79]. Finally, we discuss and conclude about the role of attention in memorability.

5.2 Memorability and eye-movement

To shed light on the relationship between images memorability and visual attention, we conducted an eye-tracking experiment on images from the memorability database proposed by [79].

5.2.1 Method

Participants and stimuli

Seventeen student volunteers (10 males, 7 females) with normal or corrected-to-normal vision took part to the eye tracking experiment. All were naïve to the purpose of the experiment and gave their full, informed consent to participate.

Class	$Avg \pm STD$	t-test
$C1$	0.82 ± 0.05	$C1$ vs $C2$, $p << 0.001$
$C2$	0.68 ± 0.04	$C2$ vs $C3$, $p << 0.001$
$C3$	0.51 ± 0.08	–

Table 5.1: Average memorability per class. The three classes are statistically different (t-test). STD represents the standard deviation.

Stimuli

We used 135 pictures extracted from [79] composed of 2222 images. We grouped them into three classes of memorability (statistically significantly different), each composed of 45 pictures. The first class consists of the most memorable pictures ($C1$, score 0.82 ± 0.05), the second of typical memorability ($C2$, score 0.68 ± 0.04) and the third of the least memorable images ($C3$, score 0.51 ± 0.08). Table 5.1 gives the main features of these classes and figure 5.1 illustrates a sample of images per class. The native resolution of the picture is 256×256. They have been resized to 384×384 to have an appropriate onscreen dimension.

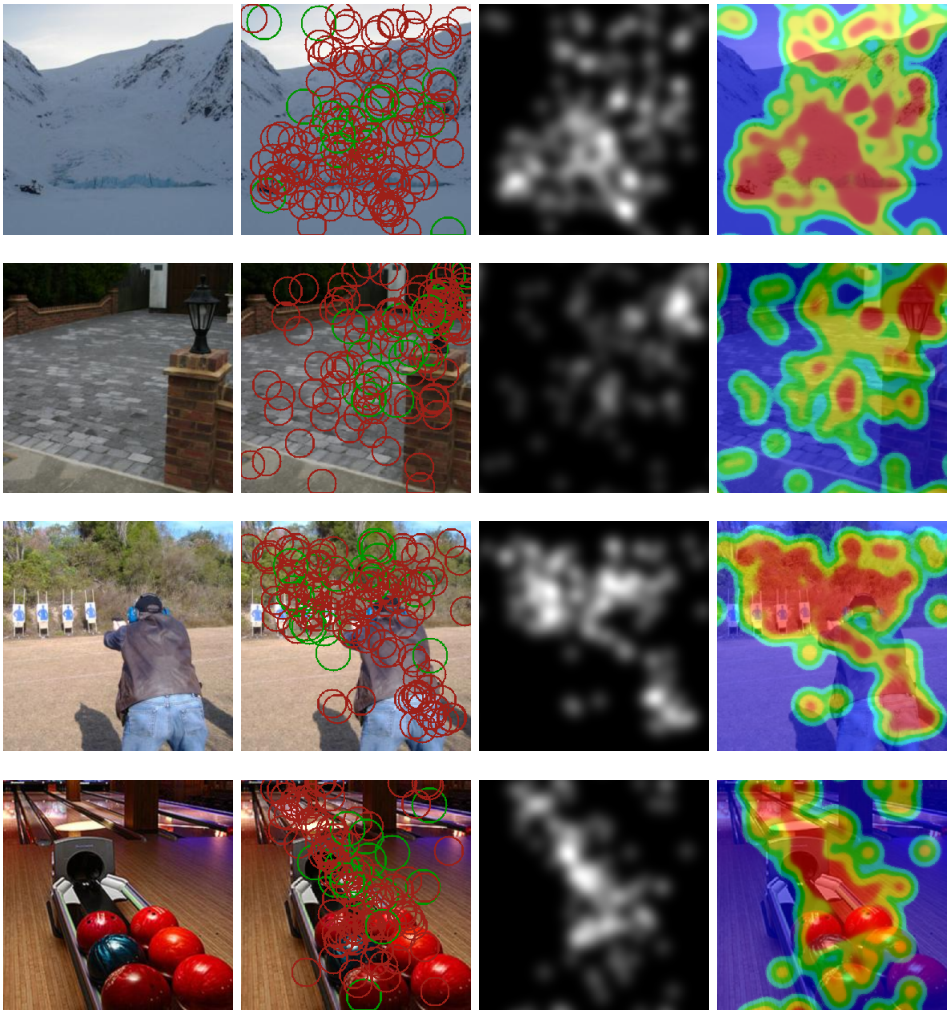
Protocol

Images were displayed on a 19 inch monitor. The square images were centered on a white background, which filled the screen resolution of 800×600 pixels. At a viewing distance of 65 cm the stimuli subtended 17 degrees of visual angle. The eyes were tracked using the Face Lab 5 eye tracking apparatus[1] with a sampling rate of 60Hz. Raw eye data were segmented into fixations and saccades by the Face Lab's system. The eye tracker is calibrated using a 9-dot grid for each participant. Three sessions, each composed of 45 pictures randomly chosen were designed. Participants were instructed to look at the pictures given that they were required to answer a question at the end of each session to ensure that they were well involved in the exercise. Each picture was displayed for 5 seconds which is enough to catch the first impression involved in memorability. Pictures were separated by a blank image displayed for 2 seconds. The participants viewed the three classes in random order to avoid any bias in the final results.

5.2.2 Results

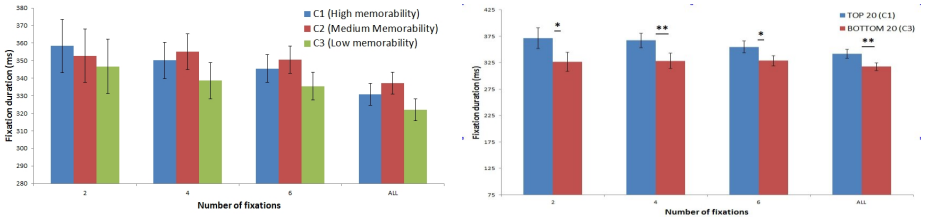
The analysis described below aims at proving that the visual behaviour of participants depends on the picture's memorability. We believe that attention is a step towards memory and therefore, this should influence the intrinsic parameters of eye movements such as the duration of visual fixations, the congruency between observers and the saccade lengths. Figure 5.1 illustrates this point. Four pictures are depicted; the first two pictures have a low memorability score whereas this score is high for the last two pictures. The first one has a memorability score of 0.81 whereas the second has a memorability score of 0.4. The average fixation durations for these two pictures are 391 and 278 ms, respectively. The average lengths of saccades are 2.39 and 2.99

[1] http://www.seeingmachines.com/product/facelab/



(a) (b) (c) (d)

Figure 5.1: (a) original pictures; (b) fixation map (a green circle represents the first fixation of observers); (c) Saliency map and (d) heat map. From top to bottom, the memorability score is 0.346, 0.346, 0.897 and 0.903, respectively (from a low to high memorability).



(a) (b)

Figure 5.2: (a) Fixation durations ($AVG \pm SEM$ (standard error of mean)) for the 3 classes function of the first 2, 4, 6 and all the fixations; (b) Fixation durations for the most and less memorable pictures. A star $*$ indicates the difference is statistically significant.

degree of visual angle, respectively. In addition, if there is something in the picture that stands out from the background, the inter-observer congruency should be higher for memorable pictures. Results are presented in the following sections. This is the case for instance for the example presented on figure 5.1.

Fixation duration

Figure 5.2 illustrates the fixation durations for the three considered classes as a function of the viewing time. The fixation durations decrease with the degree of memorability of pictures, especially just after the stimuli onset. Fixations are the longest one when observers watch memorable pictures. A statistically significant difference is found between fixation durations when the top 20 most memorable and the bottom 20 less memorable are considered. This difference is confirmed for different viewing times. These results are important since the duration of fixations reflects the deepness of the visual processing in the brain [70].

Inter-observer congruency

The congruency between observers watching the same stimulus indicates the degree of similarity between observers' fixations. A high congruency would mean that observers look at the same regions of the stimuli. Otherwise, the congruency is low. Generally the consistency between visual fixations of different participants is high just after the stimulus onset but progressively decreases over time [161]. To quantify inter-observer congruency, two metrics can be used: ROC [104] or a bounding box approach [23] (see chapter 2 section 2.3.2). The former is a parametric approach contrary to the latter. The main drawback of the bounding box approach is its high sensitivity of outliers. A value of 1 indicates a perfect similarity between observers whereas the value 0 corresponds to the minimal congruency. We used these two approaches in order to test whether memorable pictures lead to maximal inter-observer similarity (or a minimal variability). Figure 5.3 shows the congruency as a function of viewing time (only the values obtained by the ROC-based metric are given but similar results are obtained by the second method). As expected the congruency decreases over time. Results also indicate that the congruency is highest on the class $C1$ (especially after the stimuli onset (first two fixations)). The difference between congruency of class $C1$ and $C2$ is not

statistically significant. However, there is a significant difference between congruency of pictures belonging to $C1$ and $C3$. This indicates that pictures of classes $C1$ and $C2$ are composed of more salient areas which would attract more observer's attention. These results show that memorability and attention are linked. It would then be reasonable to use attention-based visual features to predict the memorability of pictures.

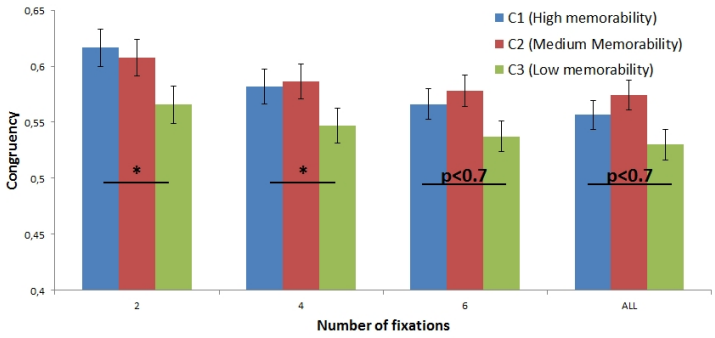
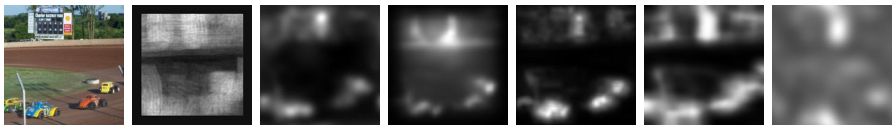


Figure 5.3: Congruency as a function of viewing time. The symbol $*$ indicates that there is a significant difference. Error bars correspond to the SEM.

5.3 Memorability prediction

Isola et al. showed that the best memorability prediction results are based on human labels containing information about the objects in the images. Nevertheless, these features are not available for any image and need time-consuming human annotations. Authors used then a mixture of several automatically extracted low-level features to approach the annotation-based results. The best result was achieved by mixing together GIST [136], SIFT [100], HOG [32], SSIM [155] and pixel histograms (PH). In this section we show that two other features of significantly smaller size which are related to attention can advantageously complement and replace some of the features proposed in [79]. For that purpose we use the SVR (Support Vector Regression) classifier and parameters from the code provided by [79].



(a) Original (b) AIM (c) AWS (d) GBVS (e) RARE (f) Seo (g) LeMeur

Figure 5.4: (a) original pictures; (b) to (g) predicted saliency map from saliency models.

5.3.1 Saliency map coverage

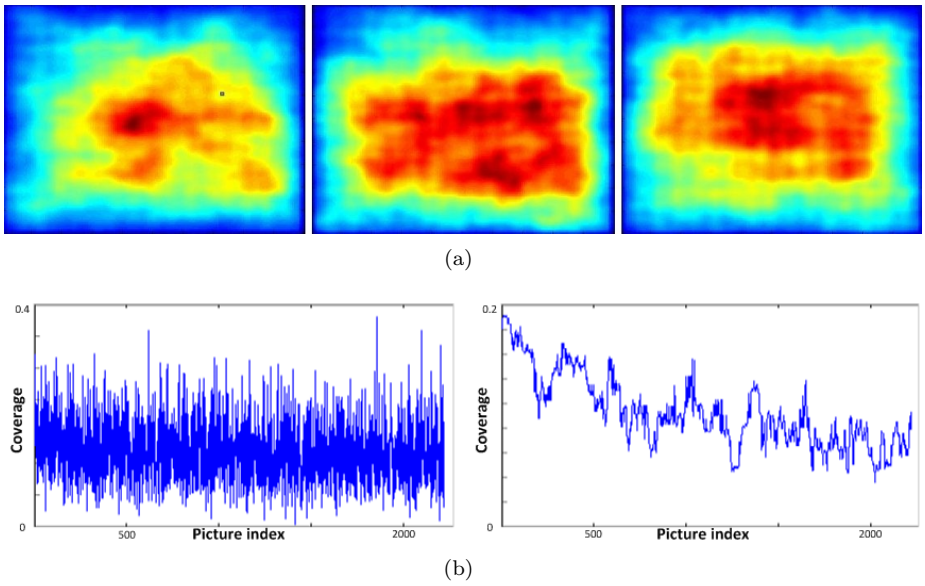


Figure 5.5: Example of average coverage using RARE algorithm: (a) for classes $C1$, $C2$ and $C3$ (from left to right) on a random collection of 45 images out of the total of 2222 images; (b) from the less to the most memorable one on the whole database. Left: unfiltered data. Right: median filtered data.

We extracted several times three classes of memorability composed of 45 images each randomly selected from a third of the most memorable images, a third of typical memorability and a third of low memorability images from the database proposed in [79]. Six state-of-the-art models of visual attention have been computed on those classes. Some saliency maps are displayed on Figure 5.4. From the saliency maps, the average saliency density is computed by accumulating the saliency maps of all the images within each class. The coverage which describes the spatial saliency density distribution is here approximated by the mean of the normalized saliency maps. A low coverage would indicate that there is at least one salient region in the image. A high coverage may indicate that there is nothing in the scene visually important as most of the pixels are attended. However, it might also indicate that there are several regions of interest which are randomly located on the images. Figure 5.5 shows the saliency coverage of the RARE [148] model. This model provides saliency maps which are the most discriminant respected to the memorability scores. We computed this saliency coverage on several randomly generated classes (and show one of them on Figure 5.5) to be sure about the result reproducibility (this result is stable independently of the chosen images). While the difference in terms of coverage between classes $C2$ and $C3$ is not obvious, this one is noticeable between the class $C1$ (the most memorable) and the two others. The class $C1$ coverage is lower which tends to show that there are mainly unique localized regions of interest while less memorable classes like $C2$ and

Figure 5.6: Low-pass filtering of images. From left to right: I_1, I_3, I_5, I_7 and I_9. RGB components are taken into account.

$C3$ either do not have precise regions of interest or have several of those regions.
The coverage of the RARE model saliency maps is thus used as a first feature in memorability prediction. Figure 5.5 (b) shows the result of the coverage for the whole database [79] from the less memorable to the most memorable image. The raw data (left plot of Figure 5.5(b)) is too noisy to be used alone (which is confirmed by the results in Table 5.2), but one can see on the median filtered version (right plot) that there is a negative correlation between the average coverage and memorability.

5.3.2 Structures (visibility)

A second feature used for memorability prediction is the contrast of the image structures. It is known [84] that object contrast is a strong attention feature. The most memorable images in Isola's database contain objects but also simpler backgrounds. This is especially true as the memorability score is established on the basis of a short observation time where complex backgrounds act like distractors and increase the visual masking phenomena.
To extract objects or at least structures contrast or 'visibility' two approaches are used together (called V1 and V2). Both are based on low-pass filtering applied several times on images with kernels of increasing sizes like in Gaussian pyramids. The kernel sizes go from 3×3 kernels which eliminate some details to 80×80 which mainly result in very fuzzy images only providing a rough idea about their context or a gist. A set of 9 images I_i with $i \in \{1, 9\}$ where the first one ($i = 1$) is the original image and the last one ($i = 9$) is the most low-pass filtered. Figure 5.6 illustrates this approach on a given picture. To quantify the impact of low-pass filtering on the images, we measure their correlation ($corr$) after filtering. In the first approach (V1) the correlations between the initial image and all the others are computed:

$$V1_i = |corr(I_1, I_i)| \qquad \forall i \in \{2, 9\} \tag{5.1}$$

In the second approach (V2) the correlation between the successively filtered images are computed:

$$V2_j = |corr(I_j, I_{j+1})| \qquad \forall j \in \{1, 8\} \tag{5.2}$$

The correlation is the mean of the correlation of the RGB components. The main idea here is to see how an image reacts to multiple low pass filtering (which might be close to the forgetting process). Contrasted strong structures will be more resistant to low-pass filtering (higher correlation) while small details and structure with cluttered background will be much less resistant and achieve lower correlation scores. Figure 5.7 shows visibility feature vectors V1 and V2 computed for the whole 2222

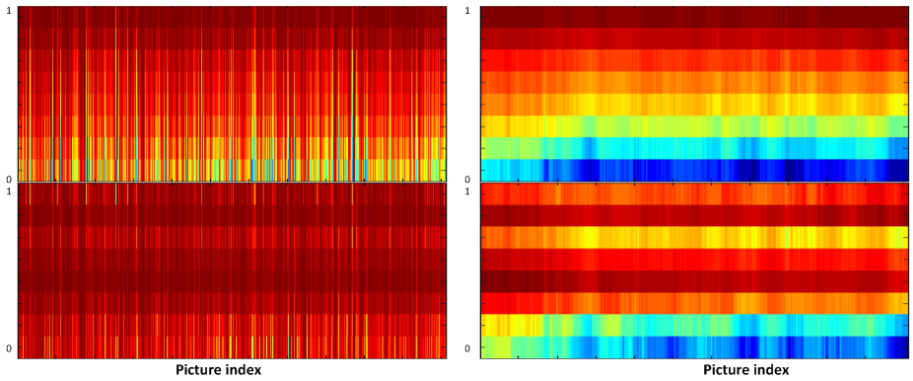


Figure 5.7: Left: raw data for the 2222 images from the less memorable to the most memorable. Right: median-filtered data. First row: V1 data, second row: V2 data.

images database. As in the previous section, the raw data both for V1 and V2 (left column of Figure 5.7) does not exhibit obvious differences. After median filtering (right column in Figure 5.7) differences between memorable and less memorable images are noticeable.

5.4 Results

The classifier and parameters are the ones from the code provided by [79]. Results shown in Table 5.2 are then perfectly comparable with the ones given in [79]. As already stated in sections 5.3.1, the proposed features are too noisy to provide good results if taken alone (see second and third column of Table 5.2). When combined to Isola et al.'s features (without the GIST which is partially redundant with the visibility low-pass filtering of our V1 and V2 features), the result is comparable and even slightly better than the one of Isola et al. (Table 5.2). The proposed attention-related features are effective when taken together with other low-level features. It should also be noted that our features perform 2% better by using 17 dimensions instead of the 512 dimensions of the GIST feature which means 86% of the total features used by Isola et al.
Table 5.3 shows the results where additional features from [79] were discarded. One by one, GIST and Pixels histograms, GIST and SIFT, GIST and HOG and GIST and SSIM were discarded from the features set. Still the results remain higher than the best combination of features in Isola et al. which shows the effectiveness of the proposed attention-related features even by replacing 1512 feature dimensions by 17 which means 58% only from the number of features in [79].

5.5 Conclusion

Isola et al. introduced an interesting approach for the prediction of image's memorability. However no relationship between attention and memorability was done. This

	Cov.	Vis.	Best (No GIST)	Best Isola
ρ	0.100	0.274	0.479	0.462

Table 5.2: Correlation results between the predicted memorability and labelled memorability. Column 2 and 3: proposed features alone (coverage, visibility). Column 4: proposed features and the SIFT, HOG, SSIM and Pixel histograms from [79], Column 5: Best feature-based combination from [79] (GIST, SIFT, HOG, SSIM, Pixel histograms).

	No Pixels	No SIFT	No HOG	No SSIM
ρ	0.476	0.474	0.470	0.468

Table 5.3: Correlation results obtained using the proposed features and combination of features excluding some features of [79] (no GIST and no Pixels histogram, no GIST and no SIFT, no GIST and no HOG, no GIST and no SSIM).

chapter shows that attention might play an important role in memorability both from an experimental and predictive perspectives when taken together with other features. The eye-tracking experiments made on a subset of the images dataset proposed by Isola et al. show that fixation duration and inter-observer congruency are well correlated with the images memorability. The prediction experiments made on the whole Isola et al. image database by using the same classifier, method and parameters showed that two attention-related features (RARE saliency map coverage and structures visibility) can advantageously replace some of the low-level features proposed in [79] and reduce in the same time the dimensionality of the feature set.

5.6 Contributions in this field

As this topic is relatively new in the community, there is only one contribution. It consists in investigating the features of eye movements when observers watch a more or less memorable pictures. Results indicate that that memorability and visual attention are linked together. An adaptation of the state-of-the-art memorability model has been done based on our behavioral conclusion.

Conference:

- M. Mancas and O. Le Meur, Memorability of natural scenes: the role of attention, ICIP, 2013.

Chapter 6

Inter-Observer Visual Congruency (IOVC)-based attractiveness. Application to image ranking

6.1 Introduction

Idiosyncrasy is defined as an individualizing quality or characteristic of a person or group, and is often used to express peculiarity (from Wikipedia). Therefore idiosyncratic eye movements refer to as the difference between the visual scanpaths of observers viewing the same stimulus. More precisely, these differences concern the intrinsic features of visual fixations. For instance, there is a strong variability of fixation durations between and within observers as shown by [145]. The causes explaining the visual dispersion are usually classified into either stimulus-dependent (or bottom-up) or observer-dependent features (or top down). Readers could find more information in Chapter 2.

In this Chapter, we present a computational model to predict the inter-observer visual congruency (IOVC). For a given picture, a score indicating the degree of visual congruence is computed. The computational model we propose, combines stimulus-dependent features which are solely inferred from the low-level visual features of the incoming picture and high-level features which are related to artistic effects for instance. We train the model by using a large eye-tracking database. In this database, we consider ee data collected during the first seconds of a picture observation.
There are very few studies dealing with the computational modeling of the inter-observer visual congruency. The closest work concerns a method to measure visual clutter which has been proposed by Rosenholtz et al. [149]. The idea is to measure the visual clutter of a scene in order to avoid confusion and to speed up the visual processing of information. For instance, a possible application is to help people to find important information on a web site or simply on a screen. Rosenholtz et al.'s solution

is based on a set of low-level visual features. The visual clutter predictor performance is assessed by comparing the amount of clutter for a scene to the reaction time required to find a target in the same scene. The proposed approach is here different since we not only use low-level visual features but also eye tracking measurements. More precisely, we use the visual scanpaths of observers in order to train a model. Our approach is supported by a number of studies suggesting that the degree of clutter present in the scene affects the deployment of our visual attention [71]. It is important to emphasize that the proposed method do not predict where people look at. It just predicts the dispersion between observers indicating whether observers look at similar locations or not.

This chapter is composed as follows. Section 6.2 gives an overview of the proposed approach. Section 6.3 describes how the IOVC is measured. A large database of eye tracking data is used for this purpose. Section 6.4 is related to the extraction of visual features that are supposed to influence the attentional allocation. Section 6.5 concerns the learning and its performance. Section 6.6 presents an application for ranking personalized pictures based on their attractiveness. Finally, we conclude the paper.

6.2 System overview

Figure 6.1 illustrates the proposed approach. First, an image database with its corresponding eye tracking data is set up. The feature extraction step extracts different visual attributes for each picture of the training dataset. Once the feature extraction is completed, the training set along with eye tracking data is used to train a cluster-weighted model. The trained model is then used to predict the inter-observer congruency of a picture taken from a new data set. Once the estimation of model's parameters has been performed, personalized photograph can be ranked according to their attractiveness. The attractiveness of an image is related to its ability to attract and to hold our attention and we assume that the attractiveness is related to the inter-observer congruency.

6.3 Measuring the inter-observer congruency

One of the key aspects of the proposed approach is to get a reliable measure of the congruency between observers. To measure this congruency, we use the method proposed by Torralba et al. [165] and described previously in Chapter 2, section 2.3.2. Just to recall, a congruency value of 1 indicates that observers fixate the same areas. Conversely, a low value would suggest that the scan patterns are uncorrelated meaning a strong variability between subjects.
To build our training data, we use Judd et al.'s dataset [92] (see Table 2.1 in Chapter 2 for more details). Figure 6.2 shows for different pictures the experimental congruency between observers. Results suggest that the congruency is small when there is nothing in the scene that catches our attention. In this case, areas that stand out the background are rare and the scene consistency is strong. When there is an object that pops out, the congruency is much higher. In addition, not surprisingly, the presence of human faces tends to increase the inter-observer congruency. It is indeed known that human faces attract in an effortlessly manner our attention. Figure 6.3 shows the distribution of the inter-observer congruency over the whole Judd's dataset. The

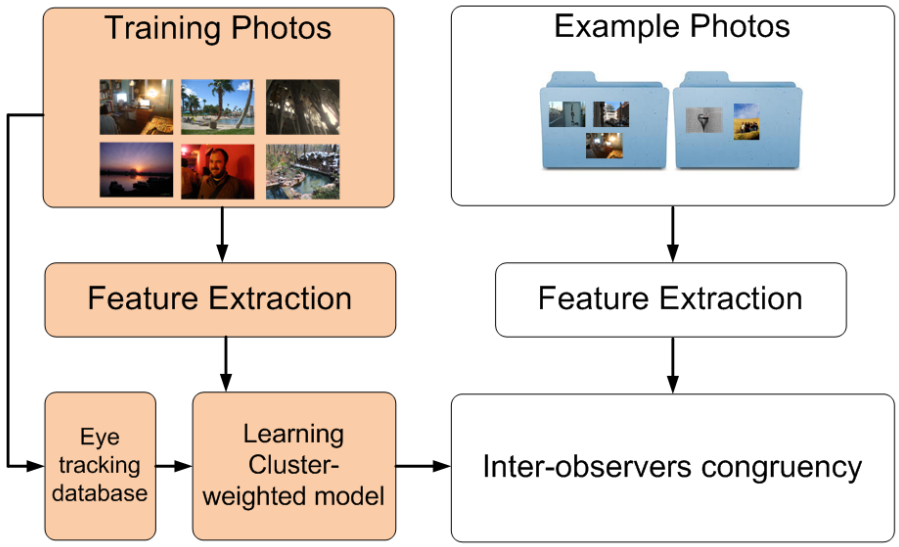


Figure 6.1: System overview.

average dispersion is of 72%, the median dispersion is of 76%. It is interesting to notice that, for a number of pictures, the congruency is maximal. This is due to the fact that a fixation point is defined by its spatial coordinates and by its neighborhood, representing one degree of visual angle (representing fovea's size).

6.4 Visual features used to predict attractivenes

In this section, the visual features used to predict the inter-observer congruency are presented. Four visual features are used. They are briefly described hereafter:

- Face detection: As the human faces significantly impact our visual deployment, it is of importance to detect human faces. The face detector we use is the one proposed by OpenCV library. The face detector is based on Haar feature-based cascade classifier for object detection. This kind of detection has been initially proposed by [173] and improved by [122].
- Color Harmony: Several studies showed that scene incongruency or inconsistency are factors influencing the inspection of an image [59, 168]. Among the scene inconsistency factors (objects, size, etc), the color might be an important factor. For instance, Frey et al. [55] showed that overt attention is significantly influenced by the presence of color. The basic assumption was that the color presence might systematically increase the congruency. The conclusion of [55] is not so straightforward. Indeed, the influence of the color might depend on the picture's category.
  The color inconsistency refers to the color harmony of the scene. We speculate that a scene with a strong consistent color harmony would be less visually disruptive than a scene with a poor color harmony. To measure the color harmony,

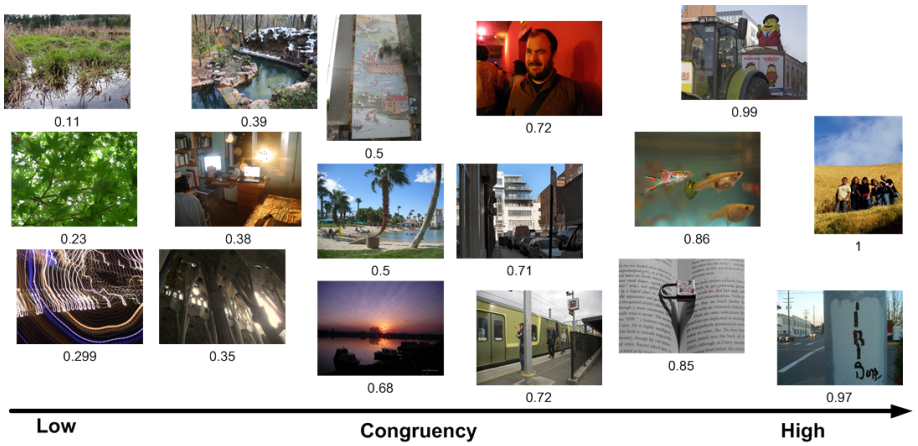


Figure 6.2: Examples of pictures associated with their corresponding inter-observer congruency. IOVC is in the range of 0 (strongest) to 1 (lowest).

we propose to follow the process of [30].

- The Depth of Field (DoF) is defined as the distance between the nearest and farthest objects in a scene that appear acceptably sharp in an image. A shallow DoF is often used to emphasize the region of interest in a picture. It is for instance used for portraiture photography. All background details are blurred whereas the nearest person (or object) is sharp, attracting our attention. An example is given figure 6.4 (a). When a large DoF is used, the opposite effect is achieved. The entire picture is sharp so that all the details of the scene are preserved. Picture of figure 6.4 (c) was taken with a large DoF.
  Estimating the DoF is then of importance. As photographers can steer our visual attention towards a particular areas by controlling the DoF, the inter-observer variability might be depend on this artistic effect.
  To determine the depth of field, the proposed algorithm relies on the fact that the shape of the horizontal/vertical derivatives histogram is modified after a blurring operation [119, 125]. The proposed scheme to compute the DoF of a picture is described below.

  Let I ($I : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$) the input picture and f_k the bluring kernel of size $k \times k$ ($k = \{3, 5, 7\}$). The blurring kernels are first applied on the luminance channel of I. then the vertical and horizontal derivatives are computed. The distributions of vertical and horizontal derivatives are given by:

$$p_{xk} \quad \alpha \quad hist(I * f_k * d_x) \tag{6.1}$$

$$p_{yk} \quad \alpha \quad hist(I * f_k * d_y) \tag{6.2}$$

  where $d_x = [1 \ -1]$ and $d_y = [1 \ -1]^T$.
  For a pixel (i, j) and for a kernel k, we compute the KL-divergence between the distributions p_{xk} and p_{yk} and the original distributions p_{x1} and p_{y1}:

$$D_k(i,j) = \sum_{(n,m) \in W_{ij}} \{KL(p_{xk}|p_{x1})(n,m) + KL(p_{yk}|p_{y1})(n,m)\} \tag{6.3}$$

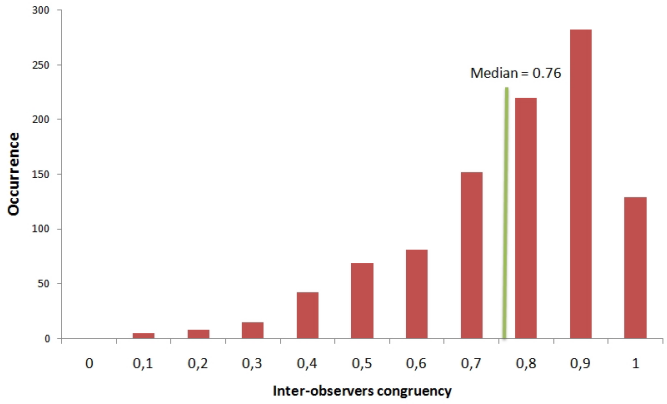


Figure 6.3: Distribution of the inter-observer congruency over Judd et al.'s dataset [92]. IOVC is in the range of 0 (strongest) to 1 (lowest).

(a) Original (b) DoF=0.12 (c) Original (d) DoF=0.75

Figure 6.4: (a) and (c) two original pictures. (b) and (d) indicates areas sensitive to blur in dark. The brigth areas correspond to unfocuss areas. DoF, standing for Depth of Field, indicates whether the picture is sensitive to blur (deep DoF) or not (shallow DoF).

where, W_{ij} is a window centered on the pixel (i, j). In this study, all the experiments were performed using an uniform kernel. The KL-divergence for a given pixel located at (i, j) is given by the following formula:

$$KL(p|q)(i,j) = p_{ij} log(\frac{p_{ij}}{q_{ij}}) \tag{6.4}$$

The KL-divergence involves two probability density functions p and q. They both sum to 1. The KL-divergence is only defined when p_{ij} and q_{ij} are greater than zero. The quantity $0log0$ is considered as zero.

The use of the KL-divergence is especially interesting in the equation 6.3 because of its similarity with the DoF values. Indeed, D_k tends to zero when the distributions p_{xk} and p_{yk} are close to p_{x1} and p_{y1}, respectively. In this case,

it means that the incoming picture is not sensitive to blur indicating that the picture is already blurred. The DoF is then low. When the value D_k increases, it suggests that the areas under analysis is rather sharp (DoF is probably high). The DoF value is finally computed as follows:

$$DoF = \sum_{(i,j)\in I} \sum_{k} D_k(i,j) \tag{6.5}$$

Figure 6.4 (b) and (d) give the value of DoF for two examples. For the first one, the DoF is of 0.12 suggesting that the picture is composed of large blurred areas. As the DoF is greater than zero the picture probably presents a sharp areas, sensitive to a blurring operation. For the second picture, the DoF is of 0.75. Unlike the previous one, this picture is more sensitive to blurring operations, suggesting that the picture is sharp. Figure 6.4 (b) and (d) illustrate in bright areas regions that are sensitive to blur. For the sake of visibility, the two pictures have been normalized in the range of 0 to 255 by using their own global maximum (3.56 and 4.68, respectively). This kind of map might be used to extract the region of interest when the DoF value is rather low, as proposed by [125].

- Visual complexity: the amount of visual information as well as the visual clutter in a picture might contribute to explain the observers' variability [149]. Oliva et al. [135] determined a list of factors that correlates with our representation of the visual complexity of a scene. Among them, the most important would be the quantity and the variety of objects, detail and color. To assess the visual complexity, three computational measures are used: the entropy, the number of regions and the amount of contours. More details are given in [105].

6.5 Learning: description and performance

6.5.1 Learning

Each image is then represented by a features vector, having a dimension of 6. The dimensionality of the features vector is not reduced as the number of dimension is low.

The estimation of the inter-observer congruency is equivalent to the estimation of the joint probability density function $p(IOVC, \mathbf{v})$. The random variable $IOVC$ represents the inter-observer visual congruency whereas $\mathbf{v}$ is the feature vector containing the six indicators. To infer the relationship between these two random variables, a learning algorithm is used. We follow the same procedure described in [163, 150] and use the software kindly provided by [150]. We just remind the main aspects of this learning procedure.

The learning consists in estimating the relationship between a measure of congruency and the extracted visual features described in the previous section. A cluster-weighted model (CWM) initially proposed by [58] is used. This is a generalization of Gaussian mixture, in which each Gaussian function expressed a part of the relationship between the input and the output distributions. The joint PDF $p(IOVC, \mathbf{v})$ is given by:

$$p_\theta(IOVC, \mathbf{v}) = \sum_{i=1}^{N} p(c_i)p(\mathbf{v}|c_i)p(IOVC|\mathbf{v}, c_i) \tag{6.6}$$

where $IOVC$ is the inter-observer congruency and $\mathbf{v}$ refers to the image features. N is the number of clusters. Each cluster is decomposed in three factors:

- $p(c_i)$ is the weight of the cluster c_i;
- $p(\mathbf{v}|c_i)$ is a multivariate Gaussian with mean μ_i and covariance matrix $\sum_i$:

$$p(\mathbf{v}|c_i) = \frac{exp\left[-\frac{1}{2}(\mathbf{v}-\mu_i)^T(\sum_i)^{-1}(\mathbf{v}-\mu_i)\right]}{(2\pi)^{L/2}\left|\sum_i\right|^{1/2}} \tag{6.7}$$

- $p(IOVC|\mathbf{v}, c_i)$ is the probability of the inter-observer congruency $IOVC$ given the input data in the cluster i:

$$p(IOVC|\mathbf{v}, c_i) = \frac{exp\left[-\frac{1}{2}(IOVC - w_i^T\mathbf{v}^*)^2\right]}{\sqrt{2\pi}\sigma_i} \tag{6.8}$$

  This is a Gaussian function with a variance equal to σ_i^2 and a mean dependent on the input feature $\mathbf{v}^*$ (same as $\mathbf{v}$ with a value 1 concatenated to its end) and a weight vector w_i. This vector indicates the weight of each input data.

The parameters θ, ($p(c_i)$, μ_i, $\sum_i$, σ_i^2, w_i, with $i = 1...N$) of the model are estimated using the Expectation-Maximization algorithm [86].

As explained in [67], in data-rich situation, it would be possible to split the data into three parts (a training set, a validation set and a test set). As this is not the case here (1000 pictures), we use the Bayesian Information Criterion (BIC) to define the model complexity. The BIC is given by:

$$BIC = -2 \times loglik + d \times \log S \tag{6.9}$$

where d is the number of free parameters depending on the number of clusters, S is the size of the dataset and $loglik$ is the maximized log-likelihood:

$$loglik = \sum_{n=1}^{S} \log p_{\widehat{\theta}}(IOVC, \mathbf{v}) \tag{6.10}$$

where $p_{\widehat{\theta}}(IOVC, \mathbf{v})$ is defined in equation 6.6. $\widehat{\theta}$ are the estimated parameters of the model.

Figure 6.5 presents the BIC values in function of N (the number of clusters). $N = 9$ is a good trade-off between complexity and quality of prediction. This value allows to predict quite efficiently the inter-observer congruency without over fitting the training data. Indeed, over fitting the data would lead to an almost perfect prediction but the risk is to loss the generalization property. As mentioned by [44], it is important to accept error to make less error. By using $N = 9$, we respect this first point. Concerning the quality of prediction, the ground truth and the predicted values of $IOVC$ are correlated $r(2004) = .34$, $p < .001$ (Pearson coefficient) and $r(2004) = .28$, $p < .001$ (Spearman coefficient).

Remark: during the learning phase, we did not use the face detector in order to limit the impact of false alarms on the estimated parameters. Instead, hand-label data are used indicating for each picture of the dataset the number of faces present.

6.5.2 Performance

A qualitative and quantitative evaluation of the proposed approach has been performed. Figure 6.6 presents some qualitative results. Ten pictures per row are given: on the top row, the first five pictures have a high IOVC whereas the last five pictures

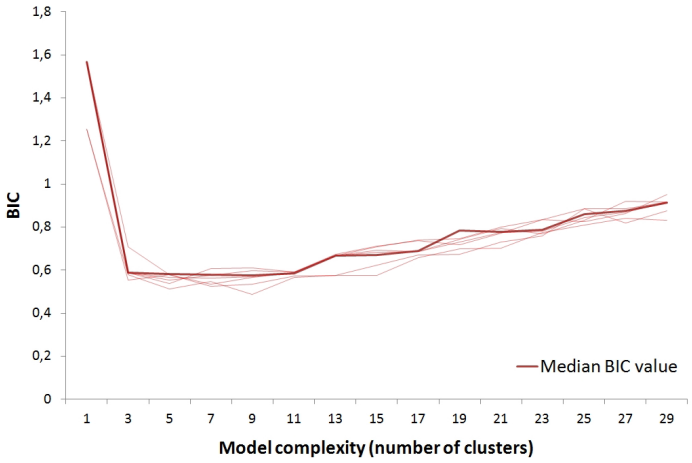


Figure 6.5: BIC in function of the model complexity. Several trials have been performed (light red curves). The dark red curve gives the median BIC values.

present a small IOVC. These results are consistent with our own subjective evaluation. The first five pictures of figure 6.6 are more attractive than the last five pictures. For these last pictures, it would be difficult to predict where an observer would look at. To illustrate this point, saliency maps of these pictures are computed using [110] (bright areas correspond to salient areas). These maps are more or less focused irrespective of the degree of attractiveness. That's why IOVC scores could be used to estimate saliency maps relevance. A high IOVC score would suggest that the saliency map has to be very focussed as for the fourth picture on the top row (Figure 6.6).

A quantitative evaluation is also performed by using another eye tracking database. This database is composed of 27 pictures. We compute the Pearson correlation coefficient between IOVC stemming from this new ground truth and our prediction. Both are correlated $r(54) = .27$, $p < .17$. The correlation is not significant due to the small number of pictures in this database. In addition, the face detector fails to detect the human faces on 5 pictures due to the varying face poses. This lack of accuracy in the detection lowers the correlation coefficient.

The proposed method is compared to the Feature Congestion measure of Rosenholtz et al. [149]. This measure aims to evaluate the visual clutter of a scene. The software available on Rosenholtz's web page is used. We run the Feature Congestion measure on the aforementioned dataset. The correlation coefficient between the Feature Congestion measure and IOVC of this dataset is $r(54) = -0.15$, $p < .43$. The correlation is negative since a high visual clutter might be interpreted as a weak congruency.

6.5.3 Limitations

The proposed model is relevant in order to predict the dispersion of observers only in free-viewing task. In the introduction, we have dressed a list of factors influencing the dispersion between observers. One factor that was not mention is the task to perform.

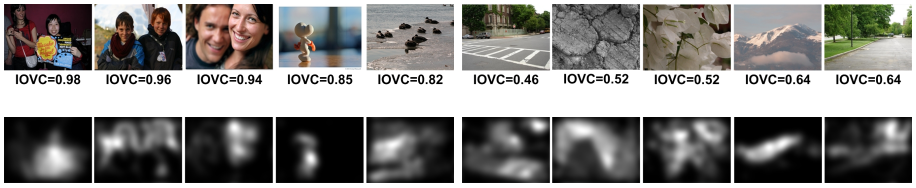


Figure 6.6: Top: pictures having high IOVC (first five) and pictures having low IOVC (last five). Bottom: saliency maps of pictures. Bright areas correspond to the most salient parts.

For instance, if we measure the inter-observer congruency when the task is to detect pedestrians, the inter-observer congruency is very high, indicating that observers share the same strategy to perform the task. To illustrate this point, we compute the inter-observer congruency over the whole eye tracking database of Ehinger [42]. The average dispersion is of 82%, the median dispersion is of 88%. Compared to the dispersion measured on Judd's database, there is a significant difference (unpaired t-test,$F(1, 1356) = 8.28$, $p < .001$).
Another limitation concerns the influence of the viewing time on the dispersion. It has been shown that the dispersion is time-dependent and increases with the time viewing. This feature is here not taken into account. For the targeted application, this feature was not judged as fundamental.
The last limitation concerns the limited accuracy of the detector we use. More specifically, as the presence of face plays an important role, the face detector has to be as efficient as possible.

6.6 Image ranking based on attractiveness

The attractiveness of an image is related to its ability to attract and to hold our attention. For instance, to give a score of attractiveness, Flickr (http://www.flickr.com) uses a combination of several parameters such as comments, annotations, favourites, etc. This is an excellent indicator but it requires a feedback or an effort of the users. An indicator based on the content analysis, such as the proposed method, might help evaluating the immediate interest of an image.

The proposed method can then be used in a context of photos browsing and automatic photograph organization. As in [151, 125, 158, 188], we propose to organize a large set of photograph. The proposed ranking is based on the picture attractiveness. This is different from state-of-the-art methods. For instance, Luo and Tang [125] ranked images according to their quality. This score is based on composition, lighting, focus controlling and color. Although there are some similarities among the extracted features (such as the DoF), better photo quality does not mean more relevant or attractiveness, as mentioned in [125]. For instance, Judd et al. [89] show that the dispersion between observers depends on image complexity and that fixations from lower-resolution images (low quality) can predict fixations on higher-resolution images (high quality).

To illustrate the proposed method, we propose to sort out forty nine images. We

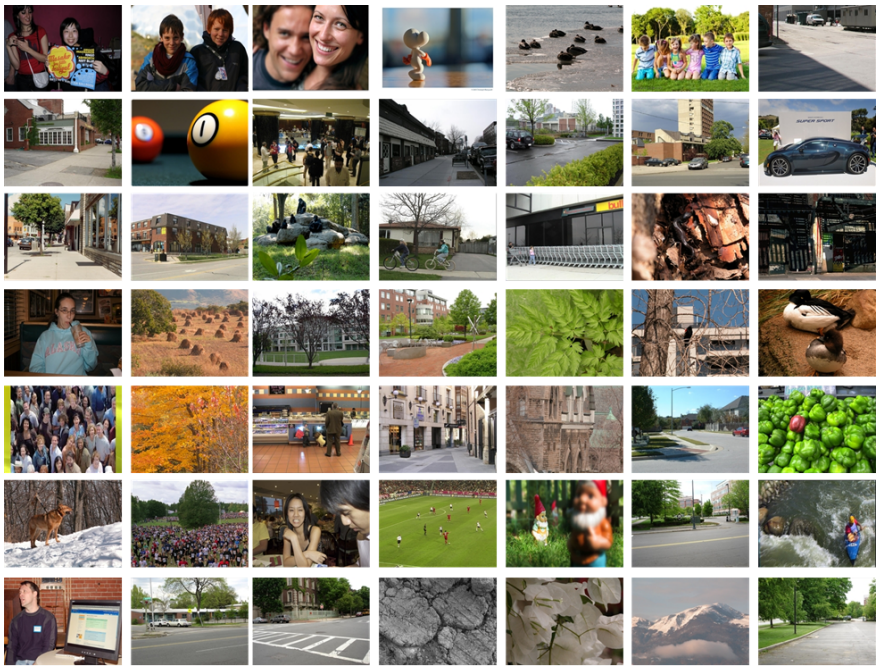


Figure 6.7: 49 pictures of various contents sorted out in function of their interestingness (from top-left (highest congruency) to bottom-right (lowest congruency)).

run the proposed model on these pictures in order to estimate their attractiveness. Figure 6.7 illustrates the results by showing the pictures ranked according to their interestingness. The first picture (top-left) has the most important IOVC whereas the picture having the lowest IOVC appears at bottom-right. On the last pictures, we can notice that there is nothing that stands from the background. In other words, it would be very difficult to predict for this kind of picture where an observer would focus on.

6.7 Conclusion

In this chapter we proposed a new criterion to automatically estimate the visual congruence between observers. We have evaluated our method qualitatively and quantitatively. We showed that our IOVC criteria outperforms the Feature Congestion measure of [149]. The predicted IOVC can be used in image processing applications where the visual perception of a picture matters such as website design, advertisement. For instance, we considered ranking personalized photograph: pictures are sorted out in function of their predicted IOVC.

However, the proposed method is still an approximation of the 'true' IOVC. It can best estimate short-term IOVC, that is the IOVC experienced in the first instant of a picture observation. In order to improve this method, it would be necessary to consider higher level factors such as those proposed by [164]. Taking into account

these factors is difficult because of their complexity.

6.8 Contribution in this field

The main contribution is the use of oculomotor data in order to carry out a learning . To the best of our knowledge, there is very few scientific contributions dealing with this point in the image processing community. This opens new avenues and perspectives for the design and improvement of image processing algorithms.
Conference:

- O. Le Meur, T. Baccino and A. Roumy, Prediction of the Inter-Observer Visual Congruency (IOVC) and application to image ranking, ACM Multimedia (long paper), 2011.

Part III

Examplar-based inpainting

Chapter 7

Examplar-based Inpainting and its variants

7.1 Introduction

Inpainting corresponds to filling holes (i.e. missing areas) in images [8]. Mathematically the problem could be formulated as described below. Let be an image $\mathbf{I}$ defined as

$$\mathbf{I} : \left| \begin{array}{l} \Omega \subset \mathbb{R}^n \longmapsto \mathbb{R}^m \\ \mathbf{x} \longmapsto \mathbf{I}(\mathbf{x}) \end{array} \right. \tag{7.1}$$

where Ω is an open set of $\mathbb{R}^n$. $n \in \mathbb{N}$ and $m \in \mathbb{N}$ are fixed integers: $n = 2$ for a 2D image and where $\mathbf{x} = (x, y)$ represents a vector indicating spatial coordinates of a pixel $p_{\mathbf{x}}$. In the case of a color image, each pixel carries three color components ($m = 3$) usually defined in the (R, G, B) color space ($I_i : \Omega \to \mathbb{R}$ represents the i^{th} image channel of $\mathbf{I}$, $i \in \{1, \ldots, m\}$). In the inpainting problem, the input image $\mathbf{I}$ is assumed to have gone through a degradation operator, denoted $\mathbf{M}$, which has removed samples from the image. As a result, the generic definition domain Ω of images can be seen as composed of two parts: $\Omega = \mathbf{S} \cup \mathbf{U}$, $\mathbf{S}$ being the known part of $\mathbf{I}$ (source region) and $\mathbf{U}$ the unknown part of $\mathbf{I}$ which we search to estimate. The degradation operator is a function $\mathbf{M} : \Omega \longmapsto \{0, 1\}$ defined as

$$M(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in U \\ 1, & \text{otherwise} \end{cases} \tag{7.2}$$

The observed degraded version $\mathbf{F}$ ($F : \Omega \longmapsto \mathbb{R}^m$) of the image can also be expressed as $\mathbf{F} = \mathbf{M} \circ \mathbf{I}$, where the $\circ$ symbol corresponds to the standard notation for the Hadamard product (pointwise multiplication). Figure 7.1 illustrates different configurations which can be encountered.

The goal of inpainting is to estimate the color components of the pixels $p_{\mathbf{x}}$ located at each position $\mathbf{x}$ in the unknown region $\mathbf{U}$. This is an ill-posed inverse problem which has no well-defined unique solution. To make this problem better defined, it is necessary to introduce image priors. The pixel values of the missing image areas are assumed to follow the same statistical or geometric structures as those in the known part of the image. These assumptions translate into different priors such as smoothness

(a) (b) (c) (d)

Figure 7.1: (a) original image; (b) the image for which 80% of the pixels have been removed; (c) the image with damaged portions in black. The purpose of inpainting is here to restore damaged portions; (d) the image where the region U (the inpainting mask) is defined by the user. We are in a context of object removal.

in a local neighborhood or sparsity. There exist a number of methods to deal with this problem. The two most important categories are briefly described below. A first class of methods encompasses variational methods, through a minimization process and diffusion-based methods using partial differential equations (PDE). This class of methods assumes some smoothness and dependence between the unknown pixels and the known part of the image in a local neighborhood. Based on this assumption, these methods smoothly propagate local image structures from the exterior to the interior of the hole. They perform well for inpainting thin (see figure 7.1 (c)) or sparsely distributed (see figure 7.1 (b)) degradations. However they are not so well adapted for texture recovery, especially when the missing region is large (see figure 7.1 (d)).

Another family of algorithms has been introduced to deal with the aforementioned limitations. These algorithms inspired by the seminal work of Efros and Leung [41] on texture synthesis rely on the assumption that statistics or structures of the textures of an image are stationary (in the case of random textures) or homogeneous (in the case of regular patterns). In other words, the known part of the image provides a good dictionary which could be used efficiently to restore the unknown part. The recovered texture is therefore learned from similar regions in a texture sample or in the known part of the image. The learning can be done simply by sampling, copying or combining pixels or patches (called examplars) from the known part of the image which are then stitched together to fill in the missing area.

A recent review [63] presents these categories in details. In this chapter we will focus specifically on the last one, also called examplar-based inpainting. This chapter is organized as follow. The seminal work of Criminisi, Pérez and Toyama [31] is first presented in section 7.2. Variants of this algorithm and our contributions regarding priority computation and hole filling, are presented in sections 7.3 and 7.4, respectively.

The following notations are used throughout this chapter:

- an image patch $\psi_{p_{\mathbf{x}}}$ is a discretized $N \times N$ neighborhood of $\mathbf{I}$ centered on the pixel $p_{\mathbf{x}}$ (N is a positive odd number). This patch can be vectorized in a raster-

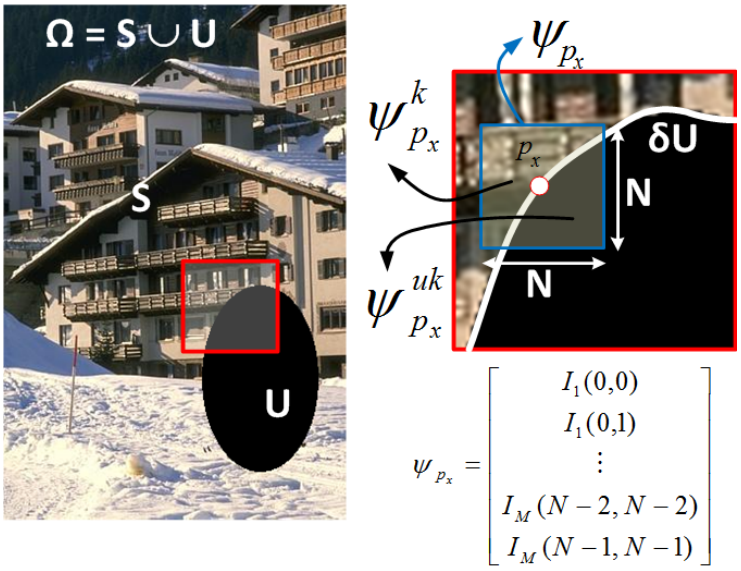


Figure 7.2: Notations used in this chapter.

scan order as a mN^2-dimensional vector (as illustrated in figure 7.2);

- $\psi^{uk}_{p_\mathbf{x}}$ denotes the unknown pixels of the patch;
- $\psi^{k}_{p_\mathbf{x}}$ denotes its known pixels;
- $\psi_{p_{\mathbf{x}(i)}}$ denotes the i^{th} nearest neighbour of $\psi_{p_\mathbf{x}}$.

7.2 Criminisi et al.'s algorithm [31]

In 2004, Criminisi et al. [31] has brought a new momentum to inpainting applications and methods. They proposed a new method based on two sequential stages which are named here filling order and texture synthesis. The first one, which is in fact the heart of the paper, is the computation of a filling order. Rather than filling the missing areas in a specific order (for instance in a raster-scan order), a priority value is computed for every pixels belonging to the front line δU; the front line is the frontier between the known and unknown areas as illustrated in figure 7.2. The priority of a given pixel $p_\mathbf{x}$ is noted $P(p_\mathbf{x})$. Once the priority has been computed for all pixels of the line front, the second stage, namely texture synthesis, begins at the spatial location having the highest priority.

7.2.1 Filling order computation

The filling order computation defines a measure of priority for each pixel of the front line. The goal is twofold: first the priority is used to distinguish areas which are easy to fill in from those which are difficult to inpaint. The idea is obviously to begin the inpainting with the simplest areas to fill. Second we would like to start the

inpainting process with the most important structures in order to propagate them into the unknown part.

The priority of a patch centered on $p_{\mathbf{x}}$ ($p_{\mathbf{x}} \in \delta U$) is composed of two terms: a confidence term C and a data term D. The priority for a pixel $p_{\mathbf{x}}$ is then given by:

$$P(p_{\mathbf{x}}) = C(p_{\mathbf{x}}) \times D(p_{\mathbf{x}}) \tag{7.3}$$

The terms C and D are described below.

Confidence term

The confidence term is the ratio of the number of known pixels divided by the total number of pixels in the patch; it then varies in the range 0 to 1 (all pixels in the patch are known). The confidence term aims to favor patches having the highest number of known pixels. This term is given by

$$C(p_{\mathbf{x}}) = \frac{\sum_{\mathbf{q} \in \psi^k_{p_{\mathbf{x}}}} C(\mathbf{q})}{\sharp \psi_{p_{\mathbf{x}}}} \tag{7.4}$$

where $\sharp$ is the number of pixels in a patch (N^2 in our case). At the first iteration, $C(\mathbf{q}) = 1 \; \forall \mathbf{q} \in S$ and 0 otherwise.

Data term

The data term is given by the absolute value of the inner product between the vector orthogonal to the gradient direction at the pixel $p_{\mathbf{x}}$, noted $\nabla I^{\perp}(p_{\mathbf{x}})$ and the vector $\mathbf{n}_{p_{\mathbf{x}}}$ which is the unit vector orthogonal to the front line δU. The data term is then given by

$$D(p_{\mathbf{x}}) = \frac{|\nabla I^{\perp}(p_{\mathbf{x}}) \cdot \mathbf{n}_{p_{\mathbf{x}}}|}{\alpha} \tag{7.5}$$

where α is a normalization constant in order to ensure that $D(p_{\mathbf{x}})$ is in the range 0 to 1.

Once the patch centered on $p_{\mathbf{x}}$ has been filled, the confidence of filled pixels is updated as follows:

$$C(p_{\mathbf{y}}) = C(p_{\mathbf{x}}) \quad \forall p_{\mathbf{y}} \in \psi_{p_{\mathbf{x}}} \bigcap U \tag{7.6}$$

This update rule implies that the confidence decreases as we move away from the front line.

7.2.2 Texture synthesis.

The filling process starts with the location $p_{\mathbf{x}^*}$ having the highest priority:

$$p_{\mathbf{x}^*} = \arg\max_{\mathbf{q} \in \delta U} P(\mathbf{q}) \tag{7.7}$$

A template matching is then performed on a window search $\mathcal{W} \subseteq S$ in order to look for the nearest neighbor $\psi^k_{p_{\mathbf{y}}}$ of $\psi^k_{p_{\mathbf{x}^*}}$ (the known pixels of the patch $\psi_{p_{\mathbf{x}^*}}$ with the highest priority):

$$p_{\mathbf{y}} = \arg\min_{\mathbf{q} \in \mathcal{W}} d(\psi^k_{p_{\mathbf{q}}}, \psi^k_{p_{\mathbf{x}^*}}) \tag{7.8}$$

where $d(a, b)$ is the sum of squared differences between patches a and b.
The pixels of the patch $\psi^{uk}_{p_{\mathbf{y}}}$ are then copied into the unknown pixels of the patch $\psi_{p_{\mathbf{x}^*}}$.

7.2.3 Some results

Figure 7.3 illustrates two inpainted pictures obtained by Criminisi's method. Results are here convincing.

(a) (b)

Figure 7.3: Inpainted pictures with Criminisi et al. method [31]: (a) input picture with a hole to be filled; (b) inpainted picture (Courtesy of P. Pérez).

7.3 Variants of filling order computation

In this section, we present two variants for computing the filling order. The first one is based on sparsity [184] whereas the second uses the structure tensor. This latter method is one of our contributions in this field [105].

7.3.1 Sparsity-based priority computation

The sparsity-based priority has been proposed recently by Xu et al. [184]. In a search window $\mathcal{W}_s$, a template matching is performed between the current patch $\psi_{p_\mathbf{x}}$ and neighboring patches $\psi_{p_\mathbf{j}}$ that belong to the known part of the image. By using a non-local means approach [180] (see section 7.4 for more details), a similarity weight $w_{p_\mathbf{x},p_\mathbf{j}}$ (i.e. proportional to the similarity between the two patches centered on $p_\mathbf{x}$ and $p_\mathbf{j}$) is computed for all patches in $\mathcal{W}_s$. These weights form the vector $\mathbf{w}_{p_\mathbf{x}}$. The sparsity

term is defined as:

$$D(p_{\mathbf{x}}) = \|\mathbf{w}_{p_{\mathbf{x}}}\|_2 \times \sqrt{\frac{|N_s(p_{\mathbf{x}})|}{|N(p_{\mathbf{x}})|}} \tag{7.9}$$

$$= \sqrt{\frac{|N_s(p_{\mathbf{x}})|}{|N(p_{\mathbf{x}})|} \times \sum_{p_{\mathbf{j}} \in \mathcal{W}_s} w^2_{p_{\mathbf{x}},p_{\mathbf{j}}}} \tag{7.10}$$

where N_s and N represent the number of valid patches (having all its pixels known) and the total number of candidates in the search window $\mathcal{W}_s$.
Given that the weights are in the range 0 to 1 and that they sum to 1, the data term is then maximal when there is only one weight equal to one. The minimal value is obtained when $w_{p_{\mathbf{x}},p_{\mathbf{j}}} = \frac{1}{|N_s(p_{\mathbf{x}})|}\ \forall p_{\mathbf{j}} \in \mathcal{W}_s$.

7.3.2 Structure tensor-based priority computation

As explained in section 7.2.1, Criminisi et al. [31] use the gradient operator in the computation of the data term. The gradient computation is relatively computational inexpensive and easy to use. However it suffers from a number of limitations. For instance, we can mention its sensitivity to noise and its limited capacity to reflect the local geometry of a scene.

To deal with such problems, we proposed to compute the structure tensor of rank 2 [178]. For a color input image ($\mathbf{I} : \Omega \to \mathbb{R}^m$) the structure tensor of $\mathbf{I}$ is the matrix function $\mathbf{J} : \Omega \to \mathbb{R}^{n\times n}$ (also called Di Zenzo matrix [34]) defined by

$$\mathbf{J} = \sum_{i=1}^{m} \nabla I_i \nabla I_i^T \tag{7.11}$$

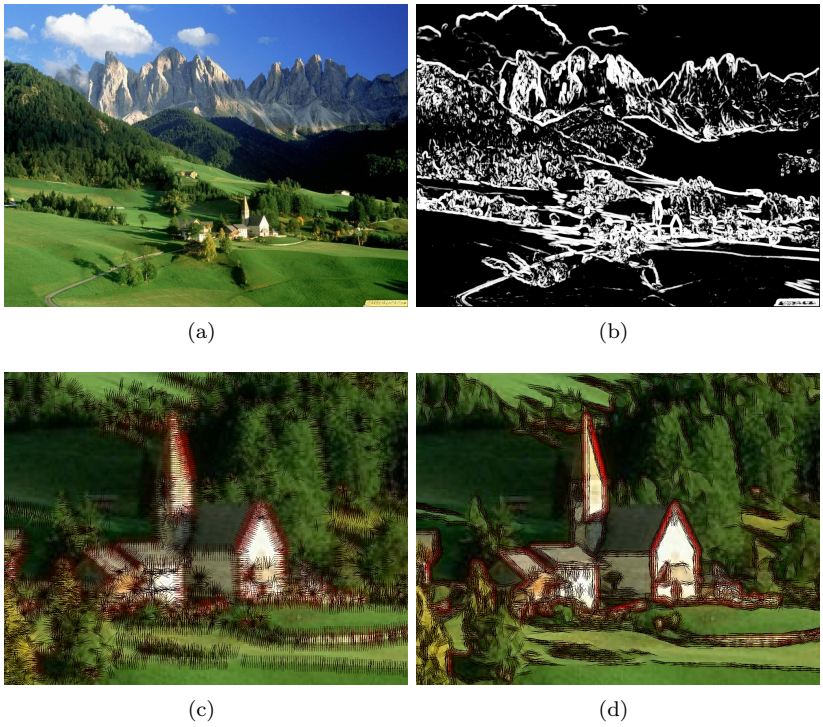
The structure tensor is a symmetric, positive semi-definite matrix. This tensor can be smoothed without cancellation effects [178]:

$$\mathbf{J}_{\rho,\sigma}[\mathbf{I}] = K_\rho * \left(\sum_{i=1}^{m} \nabla(I_i * K_\sigma) \nabla(I_i * K_\sigma)^T \right) \tag{7.12}$$

where K_a is a Gaussian kernel with a standard deviation a. The parameters ρ and σ are called integration scale and noise scale, respectively. The Gaussian filtering K_ρ serves to regularize the structure tensor field. This operation is necessary to get a smooth and regular field. However, the Gaussian kernel is isotropic and therefore might introduce severe artifacts near edges. A better solution would be to use a bilateral filtering or a non-local filter [36].

As the tensor matrix is symmetric with real coefficients, the spectral theorem indicates that there exists an orthonormal basis consisting of eigenvectors of $\mathbf{J}_{\rho,\sigma}$. An eigendecomposition is then applied to the structure tensor $\mathbf{J}_{\rho,\sigma}$ to get a precise description of the local geometry of the scene[1]. The eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$ ($\mathbf{v}_i \in \mathbb{R}^n$) define an oriented orthonormal basis and its eigenvalues $\lambda_{1,2}$ define the amount of structure variation. The tensor can then be written as $\mathbf{J}_{\rho,\sigma} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T$. From the discrepancy of the eigenvalues, the degree of anisotropy of a local region can be evaluated. Three cases can be considered:

[1] Let $T = \begin{bmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{bmatrix}$, then $\lambda_{1,2} = \frac{g_{11}+g_{22} \pm \sqrt{\Delta}}{2}$ and $\mathbf{v}_{1,2} = \begin{bmatrix} 2g_{12} \\ g_{22} - g_{11} \pm \sqrt{\Delta} \end{bmatrix}$ with $\Delta = (g_{11} - g_{22})^2 + 4g_{12}^2$.



(a) (b)

(c) (d)

Figure 7.4: (a) original image $\Omega = (0, 1024) \times (0, 768)$; (b) corresponds to the data term D. (c) and (d) illustrate the vectors $\mathbf{v}_1$ (the normal) and $\mathbf{v}_2$ (the isophote) for the central part of (a), respectively).

- if both eigenvalues are small, there is almost no variation in any direction. This is a flat region;
- if $\lambda_1 >> \lambda_2$, there is strong variation indicating the presence of an edge;
- if both eigenvalues are large, then there are variations in both directions. We are in the presence of a corner.

Concerning the two eigenvectors, the vector $\mathbf{v}_1$ indicates the orientation with the highest fluctuations (orthogonal to the image contours), whereas $\mathbf{v}_2$ gives the preferred local orientation. This eigenvector (having the smallest eigenvalue) indicates the isophote orientation. Note that when $n = 1$, $\lambda_1 = \|\nabla \mathbf{I}\|$ and $\mathbf{v}_1 = \frac{\nabla \mathbf{I}}{\|\nabla \mathbf{I}\|}$.

Similarly to [178], the data term D can then be defined according to the discrepancy of the eigenvalues as follows:

$$D(p_{\mathbf{x}}) = \alpha + (1 - \alpha) exp\left(-\frac{\eta}{(\lambda_1 - \lambda_2)^2}\right) \tag{7.13}$$

where η is a positive value and $\alpha \in [0, 1]$. When $\lambda_1 \simeq \lambda_2$, the data term tends to α. It tends to 1 when $\lambda_1 >> \lambda_2$. Figure 7.4 illustrates the data term, the vector field for the normal and isophote vectors for a given picture.

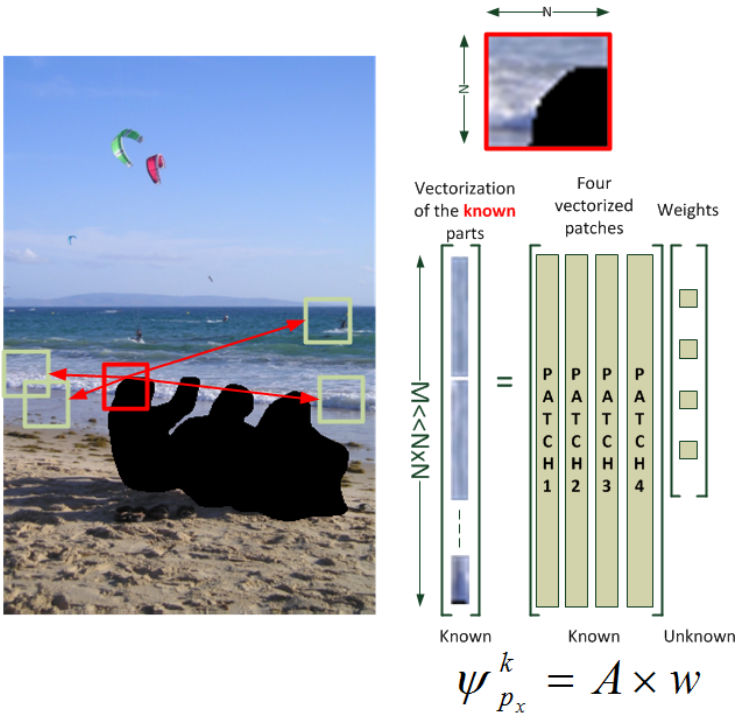


Figure 7.5: Illustration on the use of the 4 nearest neighbors. Weights are computed from the known parts of patches. They are then used to estimate the unknown parts of the patch to be filled (see formula 7.14).

7.4 Variants of texture synthesis

In Criminisi's approach, only one candidate was used to fill in the hole. A trivial extension of this work is to consider more than one candidate and to combine them. The goal is then to search for K patches $\psi_{p_{\mathbf{x}(i)}}$ where $i = 1...K$ ($\psi_{p_{\mathbf{x}(i)}} \in S$). These patches are the most similar to the known part $\psi_{p_\mathbf{x}}^k$ of the input patch. The unknown part of the patch to be filled $\widehat{\psi}_{p_\mathbf{x}}^{uk}$ is then obtained by a linear combination of the sub-patches $\psi_{p_{\mathbf{x}(i)}}^{uk}$ located at the same position as the unknown part $\psi_{p_\mathbf{x}}^{uk}$. Formally this is defined as

$$\widehat{\psi}_{p_\mathbf{x}}^{uk} = \sum_{i=1}^{K} w_i \psi_{p_{\mathbf{x}(i)}}^{uk} \tag{7.14}$$

where w_i is the weight associated to i^{th} nearest neighbor. These weights are inferred from the known parts of the patches as illustrated by figure 7.5. Methods to infer the weights are described in section 7.4.2.

The fact that we consider more than one candidate to fill in the input patch raises two main issues: the first issue is related to the number and the determination of the nearest neighbors whereas the second issue concerns the computation of weights w_i. The two next sections tackle these two points.

7.4.1 Finding the K nearest neighbours (K-NNs)

Methods and determination of K

To fill in the hole, examplar-based inpainting methods search for K nearest neighbors within the known part of the image. A naïve solution to the NN search problem is to compute the distance from the query patch to all possible candidate patches, considering each patch as a point in a high-dimensional space. Faster and approximate NN search methods exist that organize the candidates in specific space-partitioning data structures, such as the k-dimensional trees (kd-trees) [7] or the vantage point trees (vp-trees) [185]. Another solution is the generalized PatchMatch [5] which is a fast algorithm for computing dense approximate NN correspondences between patches of two images.

The number of K-NN to find is an issue. The simplest approach is to set up the K value to a fixed number of neighbours. However, a better solution is to adapt locally this number so that the similarity of chosen neighbours lies within a range $(1+\alpha) \times d_{min}$, where d_{min} is the distance between the current patch and its closest neighbour and α is a positive value. A maximum value is also used in order to limit the number of neighbours.

Similarity metrics used for generalized template matching

The similarity between two patches can be classically obtained by using a Gaussian weighted Euclidean distance:

$$d_{L^2}(\psi_{p_{\mathbf{x}}}, \psi_{p_{\mathbf{y}}}) = \|\psi_{p_{\mathbf{x}}} - \psi_{p_{\mathbf{y}}})\|^2_{2,a} \tag{7.15}$$

where a controls the decay of the Gaussian function ($g(k) = e^{-\frac{|k|}{2a^2}}$, $a \in \mathbb{R}_+$). The Gaussian weighting gives more importance to values close to the center of the patch. For a $N \times N$ patch, a reasonable choice for a is $a = \frac{N-1}{4}$. The Gaussian weights vary in $\left[e^{-4}, e^{-2}\right] \simeq [0.018, 0.05]$ [124]. This metric provides a good trade-off between matching quality and complexity. We can however improve its relevancy by considering two complementary terms such as those introduced in [22, 104]:

$$d(\psi_{p_{\mathbf{x}}}, \psi_{p_{\mathbf{y}}}) = d_{L^2}(\psi_{p_{\mathbf{x}}}, \psi_{p_{\mathbf{y}}}) \times (1 + d_H(\psi_{p_{\mathbf{x}}}, \psi_{p_{\mathbf{y}}})) \tag{7.16}$$

where $d_H(\psi_{p_{\mathbf{x}}}, \psi_{p_{\mathbf{y}}})$ is the Hellinger distance distance given by

$$d_H(\psi_{p_{\mathbf{x}}}, \psi_{p_{\mathbf{y}}}) = \sqrt{1 - \sum_k \sqrt{p_1(k) p_2(k)}} \tag{7.17}$$

where p_1 and p_2 represent the histograms of patches $\psi_{p_{\mathbf{x}}}$, $\psi_{p_{\mathbf{y}}}$, respectively. $\sum_k \sqrt{p_1(k) p_2(k)}$ is the Bhattacharyya coefficient which measures the similarity of two discrete probability distributions. d_H is null when the two distributions are strictly equal and positive otherwise. d_H is rotation and shift invariant which is not the case for d_{L^2}. This metric performs well as demonstrated in [39].

7.4.2 Inferring the weights of the linear combination

Figure 7.5 illustrates the general principles underlying the computation of the unknown part of the input patch from a set of K-NNs. The weights are first estimated from the

known parts of the K-NN. Once these weights have been determined, the same weights are applied for linearly combining the samples (see equation 7.14) corresponding to the unknown parts[2] of the K-NN.

The matrix formulation of the problem can be defined as follows

$$\psi^k_{p_{\mathbf{x}}} = \mathbf{A} \times \mathbf{w} \tag{7.18}$$

where $\psi^k_{p_{\mathbf{x}}}$ is the vectorized patch composed of the M pixels of the known part of the patch to be filled. $\mathbf{A}$ is an M-by-K matrix in which the i^{th} column represents the patch $\psi^k_{p_{\mathbf{x}(i)}}$. For the example of figure 7.5, the matrix $\mathbf{A}$ is equal to $\mathbf{A} = \left[\psi^k_{p_{\mathbf{x}(1)}} | \psi^k_{p_{\mathbf{x}(2)}} | \psi^k_{p_{\mathbf{x}(3)}} | \psi^k_{p_{\mathbf{x}(4)}}\right]$.

Average template matching

The simplest approach is to considered an uniform weighting for which all the weights are equal to $\frac{1}{K}$. This method introduces blur in the final patch. The blurry effect increases with the number of K-NNs.

Non local means

In 2006, Wong and Orchard [182] used a non local means (NLM) approach [21] to infer the weights of the linear combination. The weights are defined as

$$w_i = exp\left(-\frac{d(\psi_{p^k_{\mathbf{x}}}, \psi_{p^k_{\mathbf{x}(i)}})}{h^2}\right) \tag{7.19}$$

where h acts as a filtering parameter. The weights w_i ($i = 1..K$) depends on the similarity between the patch to be filled $\psi^k_{p_{\mathbf{x}}}$ and its i^{th} NN patch $\psi_{p^k_{\mathbf{x}(i)}}$. The weights are in the range 0 to 1 and their sum is equal to 1. The setting of the parameter h is difficult. Wexler et al. [180] define h empirically to reflect image noise.

Note that, in a denoising context, some methods have been designed to set the parameter h: some authors use a χ^2 test [93] or the Steins unbiased risk estimate (SURE) which provides the means for unbiased estimation of the true MSE (Mean Squares Error) [170].

Least squares estimation

The two previous methods do not try to minimize the approximation error which is the difference between the actual and the predicted known pixel of patches. Here, we want to solve $\psi^k_{p_{\mathbf{x}}} = \mathbf{A} \times \mathbf{w}$, i.e. we are looking for $\mathbf{w}^*$ which minimizes

$$\begin{array}{rcl} E : \mathbb{R}^K & \longrightarrow & \mathbb{R} \\ \mathbf{w} & \longmapsto & E(\mathbf{w}) = \|\psi^k_{p_{\mathbf{x}}} - \mathbf{A}\mathbf{w}\|^2 \end{array} \tag{7.20}$$

We suppose that the system is overdetermined, $K << M$. If the matrix $\mathbf{A}$ is full rank ($rank(\mathbf{A}) = K$), there is a unique solution $\mathbf{w}^*$ which is the solution of the normal equations:

$$\mathbf{w}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\psi^k_{p_{\mathbf{x}}} = \mathbf{A}^\dagger\psi^k_{p_{\mathbf{x}}} \tag{7.21}$$

[2] It is perhaps a misnomer to call that samples unknown since they are actually known. They indeed refer to samples which are collocated at the position of the unknown part of the patch to be filled.

where $\mathbf{A}^\dagger$ is the pseudoinverse of $\mathbf{A}$.

The weights obtained by the least squares method do not sum to 1 and can be positive or negative. This makes the result interpretation difficult and not invariant to translation [154].

Constrained least squares estimation, $\mathbf{w}^T\mathbf{1}_K = 1$

The sum-to-one constraint of the weight vector $\mathbf{w}$ moves the LS problem onto the locally linear embedding (LLE)[3]. The optimal weights $\mathbf{w}^*$ are determined by solving the constrained LS problem given by

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}) \qquad s.t. \qquad \mathbf{w}^T\mathbf{1}_K = 1 \tag{7.22}$$

where $\mathbf{1}_K$ denotes the K-dimensional column vector of all ones.

The problem (7.22) can be reformulated as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbf{w}^T\mathbf{B}^T\mathbf{B}\mathbf{w} \qquad s.t. \qquad \mathbf{w}^T\mathbf{1}_K = 1 \tag{7.23}$$

where $\mathbf{B}$ is a $M \times K$ neighborhood matrix. We remind that M is the number of known pixels $\psi^k_{p_\mathbf{x}}$ of the patch to be filled. The i^{th} column of $\mathbf{B}$ is equal to $\psi^k_{p_\mathbf{x}} - \psi^k_{p_{\mathbf{x}(i)}}$, where $\psi^k_{p_{\mathbf{x}(i)}}$ is the i^{th} NN of $\psi^k_{p_\mathbf{x}}$. Let $\mathbf{G} = \mathbf{B}^T\mathbf{B}$ the local $K \times K$ co-variance matrix. To minimize $E(\mathbf{w})$, we can write the Lagrangian:

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^T\mathbf{G}\mathbf{w} + \lambda\left(\mathbf{1}_K^T\mathbf{w} - 1\right) \tag{7.24}$$

By taking derivatives of $\mathcal{L}(\mathbf{w}, \lambda)$ with respect to $\mathbf{w}$ and λ and by setting them to 0, the optimal weights are the solution of

$$\mathbf{G}\mathbf{w}^* = \frac{\lambda}{2}\mathbf{1}_K \tag{7.25}$$

If the matrix $\mathbf{G}$ is invertible, the optimal weights are

$$\mathbf{w}^* = \frac{\lambda}{2}\mathbf{G}^{-1}\mathbf{1}_K \tag{7.26}$$

where λ is adjusted to ensure that everything sum to 1. In practice, we use $\mathbf{w}^* = \frac{\mathbf{G}^{-1}\mathbf{1}_K}{\mathbf{1}_K^T\mathbf{G}^{-1}\mathbf{1}_K}$.

Note that when $K > M$, (when there are more unknowns than equations), the optimization problem becomes ill-posed. A penalty term is required to stabilize the optimization. The optimal weights are in this case given by

$$\mathbf{w}^* = \frac{(\mathbf{G} + \alpha\mathbf{I}_K)^{-1}\mathbf{1}_K}{\mathbf{1}_K^T(\mathbf{G} + \alpha\mathbf{I}_K)^{-1}\mathbf{1}_K} \tag{7.27}$$

where $\mathbf{I}_K$ is the identity matrix of size $K \times K$ and α is a small constant.

This method is unfortunately sensitive to noise due to the least squares optimization. In addition, when there more unknowns than equations, a regularization is required. It involves a regularization parameter, called previously α which needs to

[3] To be more accurate, we focus on the computation of the weights $\mathbf{w}$, which is performed in the second step of the LLE algorithm.

be tuned carefully.

To solve these problems, the weights can be computed on a low-dimensional neighbourhood representation rather than using the high dimensional input space. This method is called LLE-LDNR (Low Dimensional Neighbourhood Representation). An eigendecomposition of the neighbourhood matrix $\mathbf{B}$ is given by $\mathbf{B} = \mathbf{U}\sum\mathbf{V}^T$ where $\mathbf{U} = [\mathbf{U}_1|\mathbf{U}_2]$ and $\mathbf{V} = [\mathbf{V}_1|\mathbf{V}_2]$ are orthogonal matrices of size $K \times K$ and $M \times M$, respectively. $\sum$ is a diagonal matrix of size $K \times M$ with the eigenvalues of $\mathbf{B}$ on its diagonal. The matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ contain the first r and last $K - r$ columns of $\mathbf{U}$. $\mathbf{V}_1$ and $\mathbf{V}_2$ contain the first r and $M - r$ columns of $\mathbf{V}$ respectively. The parameter r is the rank of the approximation.

The best rank-r representation of $\mathbf{B}$ is given by $\mathbf{B}_r = \mathbf{U}_1\sum_1\mathbf{V}_1^T$ where $\sum_1$ is a diagonal matrix of size $r \times r$ having in its diagonal the top r eigenvalues corresponding to the leftmost eigenvectors of $\mathbf{U}_1$. The weight vector $\mathbf{w}$ for this r-dimensional neighbourhood of the known samples of the input patch is searched in order to minimize $E(\mathbf{w}) = \mathbf{w}^T\mathbf{B}_r\mathbf{B}_r^T\mathbf{w}$. The solution is not unique and is taken as the vector in the span of $\mathbf{U}_{r+1}, ..., \mathbf{U}_K$ such that $\mathbf{w}$ has the smallest norm L_2. The optimal weights are then given by

$$\mathbf{w}^* = \frac{\mathbf{U}_2\mathbf{U}_2^T\mathbf{1}_K}{\mathbf{1}_K^T\mathbf{U}_2\mathbf{U}_2^T\mathbf{1}_K} \tag{7.28}$$

The reader could find more details in [74, 26, 64].

Constrained least squares estimation, $w_i \geq 0, \forall i$

Given that the input texture patches are non-negative, the predicted patch should also be non-negative. However the weights obtained by the previous least squares estimation can be either positive or negative weights. A natural new constraint is to impose the positivity of the weights.

The problem (7.20) becomes

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}) \qquad s.t. \qquad w_i \geq 0 \tag{7.29}$$

This problem is finally very close to a non-negative matrix factorization [114]. Two differences can be noticed:

1. we want to approximate the vector $M \times 1$ of the known samples of the patch to be filled $\psi^k_{p_\mathbf{x}}$ as the multiplication of a non-negative matrix $M \times K$ (which is here noted $\mathbf{A}$) by a vector $K \times 1$ composed of non-negative values (here noted $\mathbf{w}$): $\psi^k_{p_\mathbf{x}} \simeq \mathbf{A} \times \mathbf{w}$.
2. the second difference with a NMF framework is that the matrix $\mathbf{A}$ is known since its columns are the K-NN $\psi^k_{p_{\mathbf{x}(i)}}$ of the current patch to fill in.

Many algorithms exist to solve the non-negative matrix factorization, the most widely used being the multiplicative update approach [113]. The weights $\mathbf{w}$ are first initialized as a random dense vector with positive values. They are iteratively refined by using the following update rule:

$$\mathbf{w}^t = \mathbf{w}^{t-1} \circ \frac{\mathbf{A}^T\psi^k_{p_\mathbf{x}}}{\mathbf{A}^T\mathbf{A}\mathbf{w}^{t-1} + \epsilon} \tag{7.30}$$

where the division is a point wise division and ϵ is a small positive value to avoid the division by zero and $\circ$ is the Hadamard product.

Comparison

Table 7.1 gives a comparison test between methods previously described. The comparison aims at estimating a given patch from four neighbors. The weights are given for the different methods. The best PSNR score is given, in this example, by the constraint least squares in which the weights have to be positive or null. The average template matching performs worst. It is not surprising since this method does not optimize the approximation error.

Patch to estimate	Its first four K-NN			
				
Method	Weights		Estimated patch	PSNR (dB)
ATM	$\mathbf{w} = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right]^T$			28.45
LS	$\mathbf{w} = [-0.15, 0.79, 0.23, 0.12]^T$			30.87
LS (LLE)	$\mathbf{w} = [-0.14, 0.79, 0.23, 0.12]^T$			30.89
LS (LLE-LDNR)	$\mathbf{w} = [-0.23, 1.49, -0.19, -0.07]^T$			30.98
LS (NMF)	$\mathbf{w} = [0.00, 0.734, 0.18, 0.09]^T$			31.06

Table 7.1: Combination example: the patch to estimate is shown top-left. Its first four K-NN are given on the top row. Weights of the linear combination are estimated by five methods (ATM: Average Template Matching; LS: Least Squares; LS(LLE): Least Squares with the constraint $\sum_i w_i = 1$; LS (LLE-LDNR): Least Squares with the constraint $\sum_i w_i = 1$ and by using a low rank representation; LS (NMF)): Least Squares with the constraint $w_i \geq 0, \forall i$). The estimated patches are given as well as their quality score in terms of PSNR.

However, all these methods are sensitive to noise and outliers. It is indeed of importance to find the K-NN patches having the highest quality (i.e. the highest similarity with the patch to be filled) as possible. Unfortunately the similarity between patches is a scalar score (MSE, SAD, etc) which does not necessarily reflect the local quality of the patches. It might be an issue since the estimated weights are uniformly

applied to the whole samples belonging to patches.

7.5 Conclusion

Over the past 5 years, there has been renewed interest in the examplar-based applications and particularly in inpainting. A number of methods has then been proposed and we have just described in this chapter the most important ones (in the context of examplar-based inpainting). A more comprehensive review has recently been published in [63].

7.6 Contributions in this field

We have several contributions in this field. One of them is described in the next chapter (this contribution is not listed below). The contributions are related to the filling order computation and the combination of patches. A comprehensive survey of inpainting methods has been recently published.
Journal:

- C. Guillemot and O. Le Meur, Image inpainting: overview and recent advances, IEEE Signal Processing Magazine, 2014.
- C. Guillemot, M. Turkan, O. Le Meur and M. Ebdelli, Object removal and loss concealment using neighbor embedding methods, Elsevier Signal Processing: Image Communication, 2013.

Conference:

- O. Le Meur, J. Gautier and C. Guillemot, Examplar-based inpainting based on local geometry, ICIP, 2011.

Chapter 8

Hierarchical super-resolution-based inpainting

8.1 Introduction

The previous chapter aimed at introducing the examplar-based inpainting and its variants. Although a number of progress has been made, there are two main issues which are often overlooked. The first issue is related to the parameter settings such as the filling order and the patch size. As illustrated in figure 8.1, different settings provide different results. This is not surprising since the inpainting is an ill-posed problem. However, this sensitivity to parameters raises the question of the parameters selection: how should we define these parameters and is-it possible to find automatically a good setting? The examplar-based methods as described previously are a one-pass greedy algorithm. This is the second issue we would like to emphasize. Indeed once a patch is filled in, its value will remain unchanged until the end of the process. If an error is then performed, there is a risk to propagate it inward the hole to fill.

These two problems are addressed in this chapter by considering the combination of multiple inpainted versions of the input image. The inpainting algorithm is preferably applied on a coarse version of the input image which is particularly interesting when the hole to be filled in is large. This provides the advantage to be less demanding of computational resources and less sensitive to noise and local singularities. In this case the final full resolution inpainted image (after the combination process) is recovered by using a hierarchical Super-Resolution (SR) method. SR methods refer to the process of creating one enhanced resolution image from one or multiple input low resolution images. These problems are then referred to as single or multiple images SR, respectively. In both cases, the problem is the estimation of high frequency details which are missing in the input image(s). The proposed SR-aided inpainting method falls within the context of single-image SR. Figure 8.2 illustrates the general framework of the SR-aided inpainting.

This chapter is organized as follows. In Section 8.2, the details of the combi-

(a) Mask (b) S_1 (c) S_2 (d) S_3 (e) S_4

Figure 8.1: Inpainted pictures with different settings S_x (see [107] for the setting details). (a) original picture with the hole to be filled in. Pictures S_1 up to S_4 (from (b) to (e)) are the inpainting results (note the setting sensibility of the inpainting algorithm).

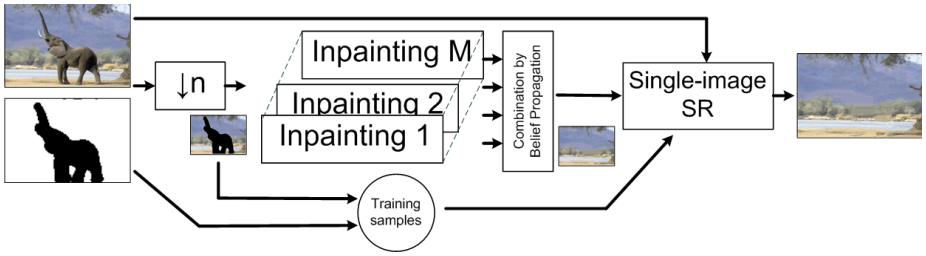


Figure 8.2: The framework of the proposed method.

nation of the inpainted pictures are given. Section 8.3 presents the super-resolution method. Experiments and comparisons with state-of-the-art algorithms are performed in Section 8.4. Finally we conclude this work in Section 8.5.

8.2 Combination of multiple examplar-based inpainting

The goal of this section is twofold: first we present the generation of multiple inpainted images. Second is related to the combination of the inpainted images.

8.2.1 Inpainting method

The proposed examplar-based method follows the two classical steps as described in the previous chapter: the filling order computation and the texture synthesis.

From the input image and a given set of parameters, we generate M inpainted image; in practise, $M = 13$. The settings we use to inpaint the image are described in [107]. A subset of them is presented in Table 8.1. Inpainted images are illustrated in figure 8.4 when these settings are used.

Table 8.1: Four settings used to fill in the unknown parts of the pictures.

Setting	Parameters
S_1 (default)	Patch's size 5×5 Decimation factor $n = 3$ Search window 80×80 Sparsity-based filling order
S_2	default + rotation by 180 degrees
S_3	default + patch's size 7×7
S_4	default + rotation by 180 degrees + patch's size 7×7

To reduce the computational complexity, the inpainting process is not applied on the full resolution images but rather on a low-resolution version of the input image.

8.2.2 Combination methods

The combination aims at producing a final inpainted image from the M inpainted images. Before delving into this subject in details, Figure 8.1 illustrates some inpainted results obtained for a given setting. We notice again that the setting plays an important role. To obtain the final inpainted picture, three kinds of combination have been considered. The first two methods are very simple since every pixel value in the final picture is achieved by either the average or the median operator as given below:

$$\widehat{\mathbf{I}}^{(*)}(p_{\mathbf{x}}) = \frac{1}{M}\sum_{i=1}^{M}\widehat{\mathbf{I}}^{(i)}(p_{\mathbf{x}}) \tag{8.1}$$

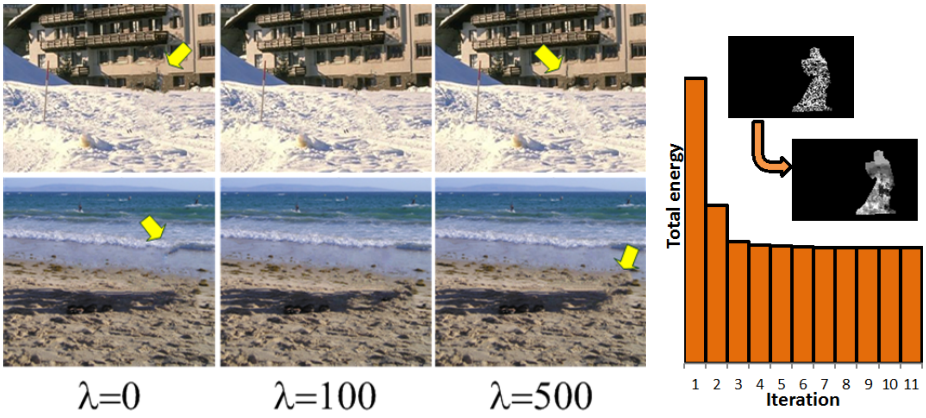
$$\widehat{\mathbf{I}}^{(*)}(p_{\mathbf{x}}) = MED_{i=1}^{M}\{\widehat{\mathbf{I}}^{(i)}(p_{\mathbf{x}})\} \tag{8.2}$$

where $\widehat{\mathbf{I}}^{(i)}$ is the inpainted picture with the i^{th} setting and $\widehat{\mathbf{I}}^{(*)}$ is the result of the combination of the inpainted pictures.
The advantage of these operators is their simplicity. However they suffer from at least two main drawbacks. The average operator as well as the median one do not consider the neighbours of the current pixel to take a decision. Results might be more spatially coherent by considering the local neighbourhood. In addition, the average operator inevitably introduces blur as illustrated by Figure 8.4.

To cope with these problems, namely blur and spatial consistency of the final result, the combination is achieved by minimizing an objective function. Rather than using a global minimization that would solve exactly the problem, we use a Loopy Belief Propagation which in practice provides a good approximation of the problem to be solved. This approach is described in the next section.

8.2.3 Loopy Belief Propagation

As in [98], the problem is to assign a label to each pixel $p_{\mathbf{x}}$ of the unknown regions U of the picture $\widehat{\mathbf{I}}^{(*)}$. The major drawback of the belief propagation is that it is



(a) (b)

Figure 8.3: Illustration of the energy minimization. (a) illustrates the influence of the smoothness term for three values of λ (see equation (8.5)); (b) the convergence process of the LBP is given for the very iterations. The labelling obtained for the first and the tenth iterations are also shown.

slow especially when the number of labels is high. Komodakis and Tziritas [98] have designed a priority Belief Propagation in order to deal with this complexity bottleneck (in [98], the number of labels is equal to the number of patches in the source region). Here the approach is simpler since the number of labels is rather small; a label is simply the index of the inpainted picture from which the patch is extracted. The finite set of labels $\mathcal{L}$ is then composed of M values ($M = 13$ here), going from 1 to M. This problem can be formalized with a Markov Random Field (MRF) $G = (\nu, \epsilon)$ defined over the target region U. The MRF nodes ν are the lattice composed of pixels inside T. Edges ϵ are the four-connected image grid graph centered around each node. We denote $\mathcal{N}_4$ this neighborhood system. The labelling assigns a label l ($l \in \mathcal{L}$) to each node/pixel $p_{\mathbf{x}}$ ($p_{\mathbf{x}} \in U$) so that the total energy E of the MRF is minimized (we denote by l_p the label of pixel $p_{\mathbf{x}}$) [17, 16]. We consider the following energy:

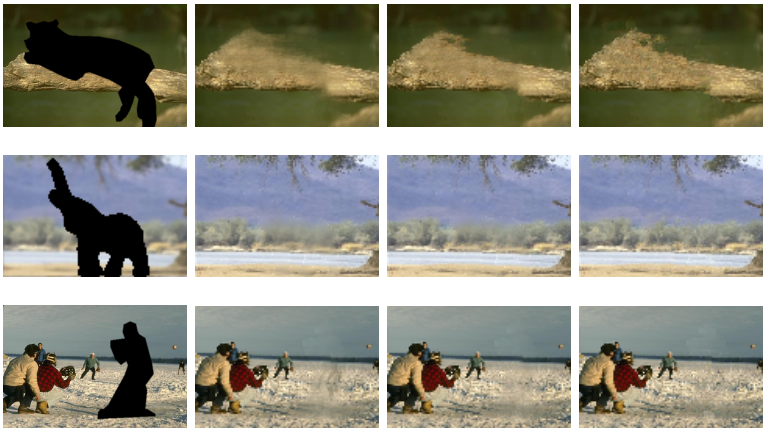
$$E(l) = \sum_{\mathbf{p} \in \nu} V_d(l_p) + \sum_{(n,m) \in \mathcal{N}_4} V_s(l_n, l_m) \tag{8.3}$$

where,

- $V_d(l_p)$ is called the label cost (or the data cost) [98]. This represents the cost of assigning a label l_p to a pixel p_x. This is given by:

$$V_d(l_p) = \sum_{n \in \mathcal{L}} \sum_{\mathbf{u} \in \upsilon} \left\{ \widehat{\mathbf{I}}^{(l)}(\mathbf{x} + \mathbf{u}) - \widehat{\mathbf{I}}^{(n)}(\mathbf{x} + \mathbf{u}) \right\}^2 \tag{8.4}$$

  where υ is a square neighbourhood (3×3) centered on the current pixel. The cost increases when the dissimilarity between the current patch and colocated patches in other inpainted pictures is high.



(a) (b) (c) (d)

Figure 8.4: Comparison of combination methods. (a) Input picture; (b) results obtained by averaging all inpainted pictures, (c) by taking the median pixel values and (d) by using a Loopy Belief Propagation.

- The pairwise potential or the discontinuity cost $V_s(l_n, l_m)$ is a quadratic cost function given by:

$$V_s(l_n, l_m) = \lambda \times (l_n - l_m)^2 \tag{8.5}$$

  where λ is a weighting factor and is set to 100. The discontinuity cost is here based on a difference between labels rather than the difference between pixel values.

The minimization of the energy E over the target region U can be achieved using loopy belief propagation (LBP) [187] and corresponds to the maximum a posteriori (MAP) estimation problem for an appropriately defined MRF [9]. Figure 8.3 illustrates the minimization of the total energy as well as the influence of the smoothness term. When $\lambda = 0$, there is no smoothness term. Some artefacts indicated by arrows are visible. When $\lambda = 500$, artefacts are visible. A good trade-off is obtained by setting the value λ to 100. On the same figure 8.3, the labelling convergence is given for four iterations (pictures on top of this figure represent label values, not pixel values). This illustrates the fact that the label choice is not greedy.

8.2.4 Comparison of the combination methods

Figure 8.4 illustrates the performance of the combination methods. As expected, when the different inpainted pictures are averaged, the reconstructed areas are blurred. The blur is less striking when the median operator is used to combine pictures. The LBP method provides the best result. The texture is well retrieved and thanks to the global energy minimization results are spatially consistent. In the following, we use the LBP method to combine low-resolution inpainted pictures.

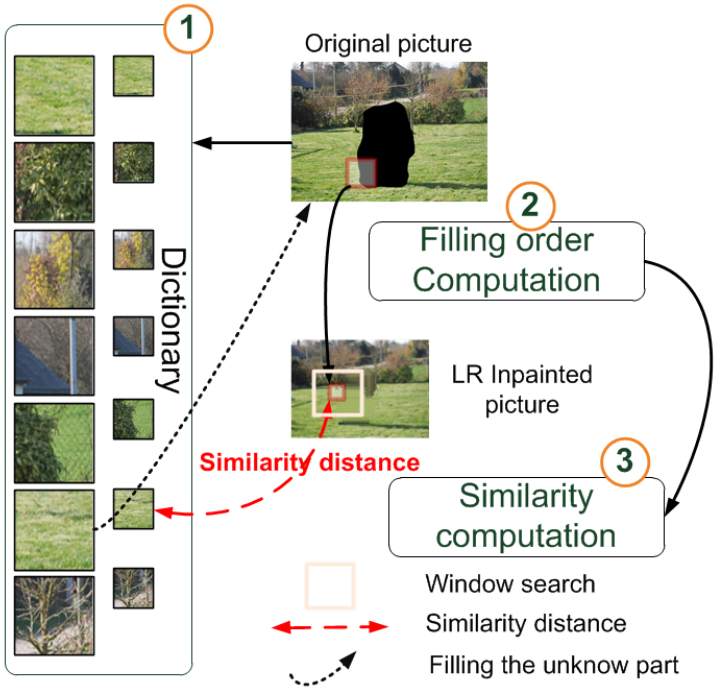


Figure 8.5: Flowchart of the super-resolution algorithm. The missing parts of the red block are filled in by the best candidate stemming either from the dictionary or from the local neighbourhood. The top image represents the original image with the missing areas whereas the bottom one is the result of the low-resolution inpainting.

8.3 Super-resolution algorithm

Once the combination of the low-resolution inpainted pictures is completed, a hierarchical single-image super-resolution approach is used to reconstruct the high resolution details of the image. We stress the point that the single-image SR method is applied only when the input picture has been downsampled for the inpainting purpose. Otherwise the SR method is not required. As in [54, 104], the problem is to find a patch of higher-resolution from a database of examples. The main steps, illustrated in figure 8.5 are described below:

1. Dictionary building: it consists of the correspondences between low and high resolution image patches. The unique constraint is that the high-resolution patches have to be valid, i.e. entirely composed of known pixels. In the proposed approach, high-resolution and valid patches are evenly extracted from the known parts of the image. The size of the dictionary is a user-parameter which might influence the overall speed/quality trade-off. Two dictionaries $\mathbf{D}^{HR}$ and $\mathbf{D}^{LR}$ are built. Their columns are the vectorized patches $\psi^{HR}_{p_\mathbf{x}}$ and $\psi^{LR}_{p_\mathbf{x}}$, respectively;
2. Filling order of the HR picture: The computation of the filling order is computed

on the HR picture with the sparsity-based method. The filling process starts with the patch $\psi_{p_{\mathbf{x}}}^{HR}$ having the highest priority and which is composed of known and unknown parts. Compared to a raster-scan filling order, it allows us to start with the linear structures and then to recover them first;

3. In the inpainted images of lower resolution, we look for the K-NN of the LR patch $\psi_{p_{\mathbf{x}}}^{LR}$ corresponding to the HR patch having the highest priority. This search is performed in the dictionary and within a local neighbourhood:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|D^{LR+}\mathbf{w} - \psi_{p_{\mathbf{x}}}^{LR}\|_2^2 \qquad s.t. \|\mathbf{w}\|_0^0 \leq L \tag{8.6}$$

where $\mathbf{D}^{LR+}$ is the dictionary $\mathbf{D}^{LR}$ extended with vectorized patches belonging to the local neighbourhood. $\|\mathbf{w}\|_0^0$ is the 'norm' L_0 which counts the number of non-zeros in $\mathbf{w}$. In the proposed approach, we choose a cardinality of 1, i.e. $L = 1$. Only the best K-NN is then chosen. We avoid the use of a linear combination of K patches which, in the context of high-resolution pictures, would introduce a blurry effect. The pixel values of the best candidate taken in $\mathbf{D}^{HR}$ are then copied into the unknown parts of the current HR patch $\psi_{p_{\mathbf{x}}}^{HR}$.

After the filling of the current patch, the priority value is propagated and the aforementioned steps are iterated while there exist unknown areas. A Poisson and alpha-blending are used to hide seams between known and unknown parts and to improve robustness.

The SR method is applied in a hierarchical manner. For instance, if the input picture of resolution (X, Y) has been down-sampled by four in both directions, the SR algorithm is applied twice: a first time to recover the resolution $(\frac{X}{2}, \frac{Y}{2})$ and a second time to recover the native resolution.

8.4 Experimental results

Some results are given in this section and illustrated by figure 8.6. A more exhaustive comparison has been performed in [107].
Figure 8.6 illustrates the results of our proposed approach compared to He [68], Shift-map results [141] and priority Belief Propagation [98] (when state-of-the-art results are available). The tested pictures are extracted from [68]. Concerning the comparison between our method and He's method, results are quite similar except for the first and last pictures for which He's approach gives more visually pleasing results, although there is no obvious artefacts in our results. Regarding shift-map and BP methods (the three last rows of figure 8.6), artefacts are visible such as on the stone wall (fourth row) and on the waterfall (BP method on the fifth row). On these pictures, our method is more robust than shift-map and Belief Propagation methods. More results are available on the following web page:
http://people.irisa.fr/Olivier.Le_Meur/publi/2013_TIP/index.html.

8.5 Conclusion

A novel inpainting approach has been proposed in 2012/2013. The input picture is first down sampled and several inpaintings are performed. The low-resolution inpainted pictures are combined by globally minimizing an energy term. Once the combination

(a) Original (b) He [68] (c) Ours

(a) Original (b) He [68] (c) Shift-Map [141] (d) BP [98] (e) Ours

Figure 8.6: Comparison with state-of-the-art results. On the three first rows: (a) original image; (b) He's results (extracted from [68]), (c) proposed approach. On the three last rows: (a) original image; (b) He's results; (c) Shift-map results [141]; (d) BP results [98] and (d) proposed approach.

is completed, a hierarchical single image-based super-resolution method is applied to recover details at the native resolution. Experimental results on a wide variety of images have demonstrated the effectiveness of the proposed method.
One interesting avenue of future work would be to extend this approach to the temporal dimension. However the main important improvement is likely the use of geometric constraint and higher-level information such as scene semantics.

8.6 Contribution in this field

The proposed method deals with two important problems of the examplar-based inpainting methods: the parameter setting and the one-pass greedy method. For that, several inpainting have been performed. Results are then combined by minimizing a global energy. The use of a super-resolution algorithm allows to work with a low-resolution version of the input picture. After the combination process, we retrieve the high-frequency by a single-image-based SR methods. Scientific publications are listed below.
Journal:

- C. Guillemot and O. Le Meur, Overview of inpainting methods, Signal Magazine Processing, 2013.
- O. Le Meur, M. Ebdelli and C. Guillemot, Hierarchical super-resolution-based inpainting and applications, IEEE Trans. On Image Processing, 2013.

Conference:

- O. Le Meur and C. Guillemot, Super-resolution-based inpainting, ECCV, pp.554-567, 2012.
- O. Le Meur, J. Gautier and C. Guillemot, Examplar-based inpainting based on local geometry, ICIP, 2011.

Part IV

Conclusion

Chapter 9

General conclusions and perspectives

This manuscript resumes my main research results since the achievement of the PhD degree. My research is cross-disciplinary spanning visual perception/cognition and image processing/editing. I focus on research questions at the intersections of these two domains. Figure 9.1 illustrates these two scientific areas where I am involved. The small bubbles indicate past and present research topics. The name of my collaborators is also indicated.

We are more and more interested in designing advanced image/video editing which would be built upon Human Visual Properties such as the visual attention. To reach this objective, it requires a deep understanding of the visual perception mechanisms, powerful models of our visual perception and advanced image editing algorithms. My perspectives on research are composed of the three following axes which are described in the following sections:

- Visual attention: the goal is to strengthen and improve my skills and knowledge in the modelling of visual attention.
- Image editing: the goal it to design and improve editing methods. Examplar-based methods are currently my main interest.
- Making the link between image editing and perceptual models.

9.1 Perspectives in the modelling of visual attention

To date, our contributions in the visual attention field cover the computational modelling of visual attention (see for instance [110], [109]), the robustness of the saliency map and methods to evaluate the similarity degree between a prediction and a ground truth [104].
As mentioned in Chapter 1, there exist a number of models which are more or less biologically plausible. These models use different mathematical tools coming from information theory, image processing or probabilistic framework. The common denominator between these models of visual attention is that they all output a 2D static

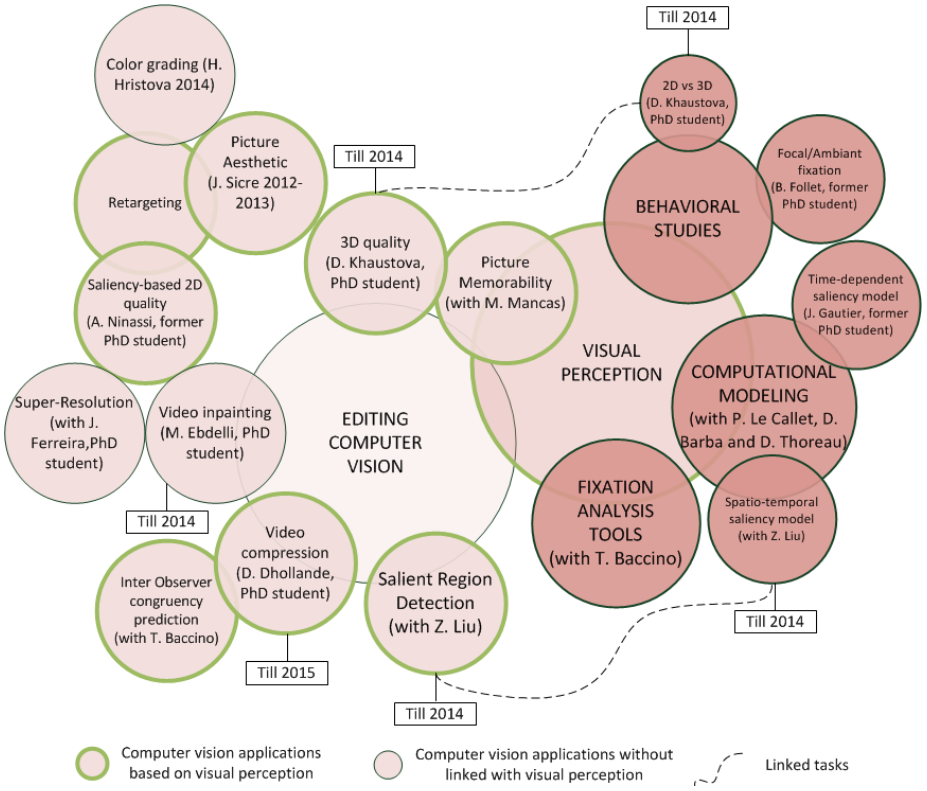


Figure 9.1: My two research areas and some contributions.

saliency map. Although that this representation is a convenient way to indicate where we look within a scene, some important aspects of our visual system are clearly overlooked. When viewing a scene, our eyes alternate between fixations and saccades, jumping from one specific location to another. This visual exploration within a visual scene is a highly dynamic process in which the time plays an important role. However most computational implementations of human visual attention could be boiled down to a simple non-dynamic map of interest. The next generation of visual attention models should be able at least to consider the temporal dimension in order to account for the complexity of our visual system and should output predicted visual scanpaths.

There are few models dealing with the generation of visual scanpaths. The first approach has been proposed by Itti et al. [84]. From a static saliency map, a scanpath is generated by using a winner-take-all (WTA) algorithm and an inhibition-of-return (IoR). Brockmann and Geisel [18] used a Lévy flight to simulate the scanpath and later Boccignone and Ferraro [11] extended Brockmann's work. Recently Wang et al. [177] used the principle of information maximization to generate scanpaths on natural images. All these previous studies suffer from the biologically plausibility and the validation of the scanpath. First the generated scanpaths should be biologically plausible and should present the same peculiarities as those of our saccadic behavior

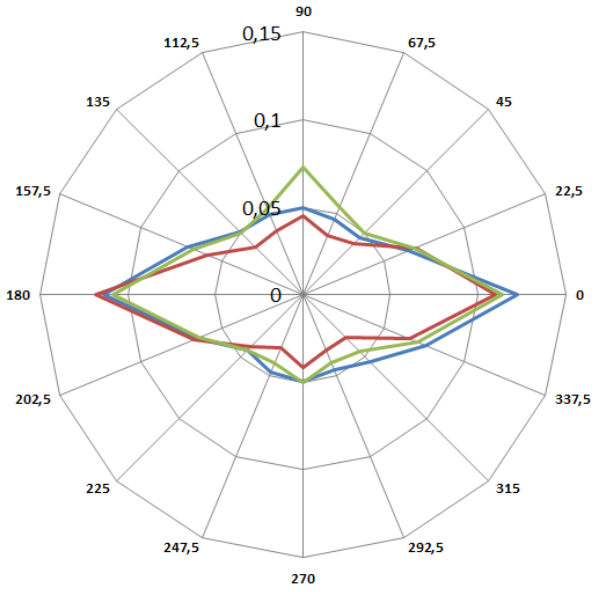
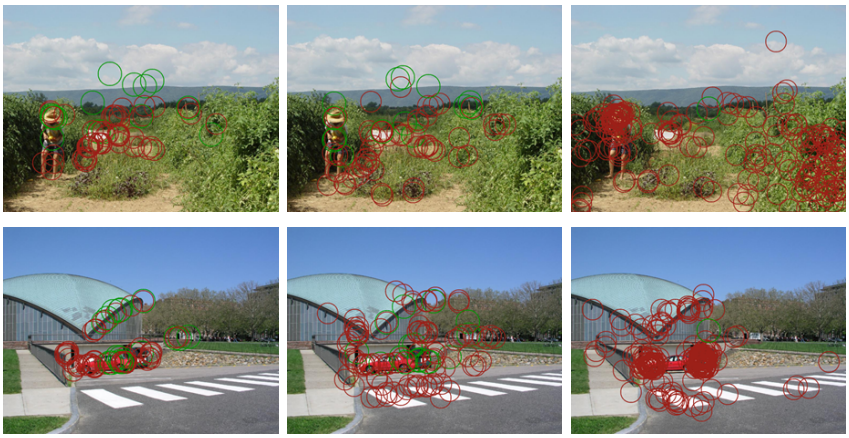


Figure 9.2: Distribution of proportion of saccades in each of the 16 orientations for 3 eye tracking data sets. All possible saccade orientations are divided into 16 bins each of 22.5 degrees. These bins are numbered in an anti-clockwise direction. The value 0 on the polar plot groups together saccades having an orientation between ± 11.25 degrees. As expected, we observe a strong horizontal bias and a proportion of oblique saccades which is much smaller than vertical and horizontal ones.

such as orientation and spatial biases. In others words, we should consider oculomotor constraints such as the geometric constraint on the length and directions of saccades to get more relevant visual scanpaths. The lack of experimental results is the second limitation of previous studies. For instance Wang et al. [177] only used 20 pictures and 3 generated scanpaths to validate their models. Finally, as these algorithms build their prediction from a saliency map, this map needs to be as accurate as possible.

We will address these problems by generating scanpaths from which static as well as dynamic saliency maps can be easily computed. To get plausible scanpaths, at least three attentional biases should be considered (some of them are presented in subsection 2.3.2 in Chapter 2): first, saccades of small amplitudes are far more numerous than long saccades. A second oculomotor bias is related to the direction of saccades: horizontal saccades are more frequent than vertical ones as illustrated by figure 9.2. Third, saccade planning is not memoryless. Several studies have shown the influence of gaze history on saccade selection [6]. Our objective is to propose a new framework for predicting visual scanpaths which present similar characteristics to those of human scanpaths. As suggested by Ellis and Smith [45], the saccade generation can be efficiently simulated by a Markov process of order T. Let $\mathcal{I}$ an input image and x_t a



(a) Our prototype (b) WTA+IoR (c) Boccignone [11]

Figure 9.3: Comparison of simulated scanpaths obtained by (a) our first prototype, (b) the WTA+IoR method (WTA=Winner Take All; IoR=Inhibition of Return) and (c) Boccignone's method [11].

fixation point at time t. To determine the next optimal fixation point, we have to consider each possible next fixation and select the location that maximizes a conditional probability given the knowledge of previous fixation locations, bottom-up saliency and oculomotor constraints. Specifically the transition between two consecutive fixation points according to the past T fixations could be defined by

$$x_t^* = \arg\max_{x \in \mathcal{I}} p\left(x | x_{t-1}, \cdots, x_{t-T}\right) \tag{9.1}$$

where x_t^* is the optimal position according to the conditional probability $p\left(x | x_{t-1}, \cdots, x_{t-T}\right)$ given by

$$\begin{aligned} p\left(x | x_{t-1}, \ldots, x_{t-T}\right) &\propto p_{BU}(x) p_M(x,t) \\ &\times p_G(d) p_\alpha(\phi) p_C(d_{sc}) \end{aligned} \tag{9.2}$$

where $p_{BU}(x)$ represents the saliency value at location x. $p_G(d)$ and $p_\alpha(\phi)$ represent oculomotor biases with respect to the amplitude d (expressed in degree of visual angle and orientation of saccades ϕ (expressed in degree) between two fixation points x_t and x_{t-1}, respectively. $p_C(d_{sc})$ is the central bias, where d_{sc} is the distance of the current location x to the screen center. $p_M(x,t)$ represents the memory state of the location x at time t. Figure 9.3 presents simulated scanpaths obtained by three methods: our first implementation of formula 9.1, a simple approach based on a winner-take-all and inhibition of return and Boccignone's method [11]. We generate 20 scanpaths, each composed of 10 fixations. This represents on one hand the common number of observers involved in an eye tracking experiment and on the other hand the common viewing time used in eye tracking experiments. The first fixation is represented by a green circle whereas the subsequent ones are represented by red circles.

Obviously, this first proposition is still very far from the reality. Indeed, we already know that there exist a number of factors, other than bottom-up ones, influencing the way we look within a scene. We can mention:

- the horizon line which unconsciously draws our attention [53].
- the contextual inferences [165] which come from the very early recognition of the scene.
- the depth cues which might play an important role. For instance, for urban scenes, the perspective depth cues, vanishing lines and points tend to attract our gaze. For synthetic stimuli, the orientation of the surfaces in depth might also hypothetically direct the saccades, as observed by Wexler et al. [179].

In addition, most of existing methods predict where fixations are directed but they do not account for fixation durations. The average fixation duration during scene viewing is about 300 ms. However, as discussed in Chapter 2 (see section 2.3), there exist a substantial variability around this mean both within an individual and across individual [145]. The fixation durations depend in fact on a number of visual and cognitive factors associated with the currently fixated region. Nuthmann et al. [133] termed this kind of fixation as being under *direct control*. A future investigation will be to predict how long fixations remain in a given location.

To conclude, this research theme aims to design new computational models of visual attention taken into accounting attentional biases and variations in fixation durations. To complete successfully these tasks, we will have to consider several open issues which are more and more discussed in the community [147, 15]. The most important ones are briefly detailed below:

- Data set: the first data sets which were used to evaluate the prediction quality of saliency models are unfortunately limited and contaminated by different biases. Therefore there is a need to revisit eye tracking protocol in order to build the best representative data set. For that, we need to identify and to understand the weaknesses of current ground truths. The current trends consist in evaluating the influence of large, medium and small salient regions on the prediction quality of saliency models, in measuring the dispersion between observers, etc. Another aspect is to provide a data set which is composed of both manually segmented salient regions (human labelled) and eye fixation ground truth (collected with an eye tracker). The first one has been proposed by [120]. This data set contains 235 color images and can be divided into six different groups of images.
- Center-Bias: people working on the modelling of visual attention have to deal with this critical issue. This tendency to look towards the center of an image (see chapter 2 for more details) needs to be considered carefully. Indeed, several studies have shown that a trivial model which predicts salience areas near the center outperforms much more complex saliency models. The most simple model is a simple Gaussian function centered on the screen. For this reason, experimenters have to envision new solutions and protocols to lessen it as much as possible.
- Metrics: as described in Chapter 3, there exist a number of similarity metrics. They compare either saliency maps, fixation points or both. These scores have serious disadvantages which unfortunately outweigh their benefits and can result sometimes in misinterpretation. For instance, the AUC can be extremely high regardless of the false alarm rate [190]. A high AUC can be observed as soon as the hit rate is high. Other issues of existing metrics are their sensitivity to transformation (for instance peak-to-peak normalization) or to smoothing (for instance when a Gaussian weighting is applied). There is still a need to define

a new metric which would be invariant to transformation, to center-bias and would have well defined bounds.

- Top-down influences: to improve upon the quality of the saliency prediction, we have to consider the use of high-level information. Some models already embed simple top-down features such as faces [92], people [92], cars [92], text [24] and horizon line [102]. However, the ideal would be to infer quickly the type of the scene and then to benefit from this prior knowledge to adapt and modulate the computation of saliency map.

9.2 Perspectives in image editing

In this manuscript, we presented examplar-based inpainting for still color pictures. In the future we will pursue our research on this topic. Two particular points are targeted: video inpainting and quality assessment of inpainted images.

The extension of inpainting to video sequences is the main target. M. Ebdelli (PhD student under the supervision of C. Guillemot and myself) is currently working on this task. The goal is to remove an object from a complex dynamic scene. In the preliminary version of our video inpainting algorithm, the first step is to perform the alignment of a set of frames with respect to the current frame. This registration is achieved by a new region-based homography computation. Once the frames have been aligned, they form a stack of images from which the best candidate pixels are searched in order to replace the missing ones. The best candidate pixel is found by minimizing globally a cost function. The first results are promising.
There are however various avenues of improvement. Indeed most of video inpainting algorithms make several assumptions which are difficult to remove. For instance, two binary masks are required to perform the video inpainting in the method proposed by Granados et al. [60]. One mask classically indicates the object we want to remove from the scene. This mask needs to be as accurate as possible especially on the region boundaries. A second mask is used to indicate the spatial positions of foreground objects. The goal is here twofold. First is to protect these areas from the inpainting process. Second is to avoid the propagation of foreground textures into the background. The building of these masks on a frame basis is difficult and time-consuming, making it hard for people to use it. Another issue is the spatio-temporal consistency over the video sequence. To deal with this point, one solution is to perform the inpainting on the whole video sequence as in [60]. However, this solution is time-consuming and very limited in term of applications. The challenge, therefore, is to develop an inpainting method that is both flexible and realistic, and for which the user interaction is as small as possible.

A second avenue related to inpainting is about the definition of an objective quality metric. As the inpainting problem is an inverse problem, there exist more than one solution. Estimating the quality of the reconstruction is then not easy. First there is no reference that can be used to make a comparison. Second the quality term is of special significance when we are talking about inpainting. We have to assess not only the signal quality (such as in conventional quality metric) but also and more importantly the structure and the coherency of the inpainted areas with respect to the known part of the scene. The former should take into account for instance the blur introduced by inpainting method, the modification of the contrast and the color coherency of the

(a) Original image (b) Style to use (c) Result

Figure 9.4: Our preliminary results on color grading inspired by [12]. (a) original image; (b) style to transfer into the original image; (c) result. Courtesy of H. Hristova.

inpainted area. The latter point is related to the structure continuity as well as its relevance. This aspect may be more difficult to evaluate faithfully. To the best of our knowledge there is only one recent paper dealing with this problem. Dang et al. [33] evaluate the similarity degree of structure and hue which exists between inpainted and known parts of the scene. In addition they propose to detect salient structure; they assume that the contours and other relevant structures in the inpainted regions attract more human gaze than the other components. This method is the first proposition for the quality assessment of inpainted areas. Although based on relevant principles, this method suffers from several issues. First the assumption that edges are more salient than other areas is questionable. Second the proposed metric is a parametric method but the parameters sensitivity has not been evaluated. Third the evaluation of the metric's relevance has been performed with a small data set (only 6 pictures). Last but not the least the structure similarity used by [33] is performed at a unique scale which does not allow to detect a large variety of inpainting artefacts.

In addition to the two aforementioned points, we have already started the exploration of new and different avenues regarding examplar-based applications. They encompass a number of methods whether they be colorization, color transfer/grading, aesthetic style transfer or super-resolution. Figure 9.4 illustrates our first result in color grading. The idea is to modify an input image/video sequence according to a color grading style specified by an user. Regarding single-image super-resolution, figure 9.5 illustrates a new method based on the use of structure tensor. Structure tensor is used to define a streamline orthogonal to salient edges. By filtering the pixel values along the streamline, it is possible to sharpen edges.

Our objective is to design image editing algorithms which could be easily modified and upgraded by using perceptual properties of the human visual system.

9.3 Perceptual-based editing

Turn on your TV set, open a magazine or walk around in a street, our visual system has to deal with a large amount of visual data. To reduce visual information to process, we select the most important visually information according to either a bottom-up or top-down mechanisms. However, have you paid attention to the print, TV broadcast or billboard advertising space? Even if you skim a newspaper without paying too

(a) Original image (b) Nearest interpolation (c) Dong's method [35] (d) Proposed (e) Zoom in (c) (f) Zoom in (d)

Figure 9.5: Our preliminary result on super-resolution. (a) original image; (b) Nearest interpolation with a magnification factor of 4; (c) Dong's method [35]; (d) proposed method based on structure tensor. Courtesy of J.C. Ferreira.

much attention, it appears fundamental for an advertiser that his advertisement attracts your attention effortlessly and unconsciously. For that, the bottom-up saliency should be maximal over areas connected to the advertisement message.

The final goal of our research is to combine the modelling of visual attention with image/video editing methods. More specifically it aims at altering images/video sequences in order to attract viewers attention over specific areas of the visual scene. More specifically, the goal of perceptual-based editing methods we intend to design is to provide a computational editing model which emphasizes and optimizes the importance of pre-defined areas of the input image/video sequence. There exist only few attempts in the literature dealing with this problem [183], [159]. These methods simply alter the content by using a blurring operation or by recoloring the image so that the focus of attention falls within the pre-defined areas of interest. We intend to go further by minimizing a distance computed between a user's defined visual scanpath and predicted visual scanpath. Iteratively the content is edited (i.e. recoloring, region rescaling, local contrast/resolution adjustment, removing disturbing object, etc) in order to move the focus of attention towards those selected by the user.

Bibliography

[1] A. Acik, A. Sarwary, R. Schultze-Kraft, S. Onat, and P. Konig. Developmental changes in natural viewing behavior: Bottom-up and top-down differences between children, young adults and older adults. *Frontiers in Psychology*, 1, 2010. 28

[2] H. Alers, L. Bos, and I. Heynderickx. How the task of evaluating quality influence viewing behavior. In *QoMEX*, 2011. 32

[3] Q. Amatya, N. Gong and P.C. Knox. Differing proportions of express saccade makers in different human populations. *Exp Brain Res.*, pages 117–129, 2011. 28

[4] T. Baccino. *La Lecture lectronique [Digital Reading].* Grenoble : Presses Universitaires de Grenoble, Coll. Sciences et Technologies de la Connaissance, 2004. 45

[5] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009. 95

[6] P. M. Bays and M. Husain. Active inhibition and memory promote exploration and search of natural scenes. *Journal of Vision*, 12(8):1–18, 2012. 113

[7] J. Bentley. Multidimensional binary search trees used for associative searching. In *Commun. ACM*, volume 18, pages 509–517, 1975. 95

[8] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *SIGGRPAH 2000*, 2000. 87

[9] A. Blake and A. Zisserman. *Visual reconstruction.* MIT Press, 1987. 105

[10] P. Blignaut. Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, 71(4):881–895, 2009. 24

[11] G. Boccignone and M. Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(12):207 – 218, 2004. 112, 114

[12] N. Bonneel, K. Sunkavalli, S. Paris, and H. Pfister. Example-based video color grading. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2013)*, 32(4), 2013. 117

[13] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013. 6, 7, 17, 30

[14] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, pages 1–16, 2012. 40, 42, 56, 57

[15] A. Borji, H.R. Tavakoli, D.N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency modeling. In *ICCV*, 2013. 115

[16] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. On PAMI*, 26(9):1124–1137, 2004. 104

[17] Y. Boykov, O. Veksler, and R.Zabih. Efficient approximate energy minimization via graph cuts. *IEEE Trans. On PAMI*, 20(12):1222–1239, 2001. 104

[18] D. Brockmann and T. Geisel. The ecology of gaze shifts. *Neurocomputing*, 32(1):643–650, 2000. 112

[19] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems*, volume 18, pages 155–162, 2006. 32, 38, 56

[20] N.D.B. Bruce and J.K. Tsotsos. Saliency, attention and visual search: an information theoretic approach. *Journal of Vision*, 9:1–24, 2009. 8, 14, 16

[21] A. Buades, B. Coll, and J.M. Morel. A non local algorithm for image denoising. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 60–65, 2005. 96

[22] A. Bugeau, M. Bertalmío, V. Caselles, and G. Sapiro. A comprehensive framework for image inpainting. *IEEE Trans. on Image Processing*, 19(10):2634–2644, 2010. 95

[23] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46(26):4333–4345, Dec 2006. 36, 69

[24] M. Cerf, J. Harel, W. Einhauser, and J. Koch. Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems*, volume 20, pages 241–248, 2007. 32, 116

[25] C. Chamaret, O. Le Meur, and J.C. Chevet. Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies. In *ICIP*, pages 1077–1080, 2010. 10

[26] J. Chen and Y. Liu. Locally linear embedding: a survey. *Artif. Intell. Rev.*, 36:29–48, 2011. 98

[27] S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: A bayesian inference theory of visual attention. *Vision Research*, 55:2233–2247, 2010. 32

[28] H.F. Chua, J.E. Boland, and R.E. Nisbett. Cultural variation in eye movements during scene perception. In *Proceedings of the National Academy of Sciences*, volume 102, pages 12629–12633, 2005. 28

[29] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Erlbaum Associates, 1988. 29

[30] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.Q. Xu. Color harmonization. In *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, volume 56, pages 624–630, 2006. 78

[31] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by examplar-based image inpainting. *IEEE Trans. On Image Processing*, 13:1200–1212, 2004. iii, 88, 89, 91, 92

[32] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 70

[33] T. T. Dang, B.A. Beghdadi, and C.C. Larabi. Perceptual quality assessment for color image inpainting. In *ICIP*, 2013. 117

[34] S. Di Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33:116–125, 1986. 92

[35] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Trans. On Image Processing*, 20(7):183857, 2011. 118

[36] V. Doré, R. F. Moghaddam, and M. Cheriet. Non-local adaptive structure tensors. *Image and Vision Computing*, 29(11):730–743, 2011. 92

[37] Michael Dorr, Thomas Martinetz, Karl R Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10), 2010. 32

[38] M.W.G. Dye, C.S. Green, and D. Bavelier. The development of attention skills in action video game players. *Neuropsychologia*, 47(8-9):17801789, 2009. 30

[39] M. Ebdelli, O. Le Meur, and C. Guillemot. Analysis of patch-based similarity metrics: Application to denoising. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 2070 – 2074, 2013. 95

[40] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. New-York: Chapman and Hall., 1993. 53

[41] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1033–1038, 1999. 88

[42] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes. *Visual Cognition*, 17:945–978, 2009. 32, 83

[43] W. Einhauser and P. Konig. Does luminance-contrast contribute to a saliency for overt visual attention? *European Journal of Neuroscience*, 17:1089–1097, 2003. 53

[44] H.J. Einhorn. Accepting erro to make less error. *Journal of Personality Assessment*, 50(3):387–395, 1986. 81

[45] S. R. Ellis and J. D. Smith. *Patterns of statistical dependency in visual scanning*, volume 9, pages 221–238. Elsevier Science Publishers BV, Amsterdam, 1985. 113

[46] U. Engelke, A.J. Maeder, and H.J. Zepernick. Visual attention modeling for subjective image quality databases. In *MMSP*, 2009. 32

[47] K. Evans, C. Rotello, X. Li, and K. Rayner. Scene perception and memory revealed by eye movements and receiver - operating characteristic analyses: Does a cultural difference truly exist? *Q J Exp Psychol*, 62:276–285, 2009. 28

[48] Facebook, Ericsson, and Qualcomm. A focus on efficiency. Technical report, Facebook, 2013. 1

[49] J.H. Fecteau and D.P. Mumoz. Salience, relevance and firring: a priority map for target selection. *Trends Cogn Sci*, 10(8):382–390, 2006. 26

[50] J. M. Findlay. Saccade target selection during visual search. *Vision Research*, 37:617–631, 1997. 1

[51] B. Fischer, H. Weber, M. Biscaldi, F. Aiple, P. Otto, and V. Stuhr. Separate populations of visually guided saccades in humans: reaction times and amplitudes. *Exp Brain Res*, 92:528541, 1993. 28

[52] C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. John Wiley and Sons, 2011. 35

[53] T. Foulsham, A. Kingstone, and G. Underwood. Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research*, 48:1777–1790, 2008. 12, 13, 15, 26, 37, 115

[54] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002. 106

[55] H.P. Frey, C. Honey, and P. Konig. What's color got to do with it? the influence of color on visual attention in different categories. *Journal of Vision*, 8(14), October 2008. 77

[56] D. Gao and N. Vasconcelos. Bottom-up saliency is a discrimant process. In *ICCV*, 2009. 8

[57] J. Gautier and O. Le Meur. A time-dependent saliency model mixing center and depth bias for 2d and 3d viewing conditions. *Cognitive Computation*, 4(2):141–156, 2012. 13

[58] Gershnfel. *The nature of mathematical modelling*. Cambridge, Univ. Press, 1999. 80

[59] R.D. Gordon. Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 30:760–777, 2004. 77

[60] M. Granados, J. Tompkin, K. Kim, O. Grau, J. Kautz, and C. Theobalt. Background inpainting for videos with dynamic objects and a free-moving camera. In *European Conference on Computer Vision (ECCV)*, 2012. 116

[61] C.S. Green and D. Bavelier. Action video game modifies visual selective attention. *Nature*, 423(6939):534537, 2003. 30

[62] D. Green and J. Swets. *Signal detection theory and psychophysics*. New York: John Wiley, 1966. 50

[63] C. Guillemot and O. Le Meur. Image inpainting: Overview and recent advances. *IEEE Signal Processing Magzine*, 2014. 88, 100

[64] C. Guillemot, M. Turkan, O. Le Meur, and M. Ebdelli. Object removal and loss concealment using neigbor embedding methods. *Signal processing: image communication*, 28:1405–1419, 2013. 98

[65] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Computer Vision and Pattern Recognition*, 2008. 9

[66] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its application in image and video compression. *Trans. On Image Processing*, 19(1):185–198, 2010. 9

[67] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics, 2001. 81

[68] K. He and J. Sun. Statistics of patch offsets for image completion. In *ECCV*, 2012. 107, 108

[69] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498 – 504, 2003. 27

[70] J.M. Henderson. Regarding scenes. *Current Directions in Psychological Science*, 16:219–222, 2007. 33, 62, 69

[71] J.M. Henderson, M. Chanceaux, and T.J. Smith. The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, 9(1), January 2009. 76

[72] T. Ho-Phuoc, A. Guerin-Dugue, and N. Guyader. A biologically-inspired visual saliency model to test different strategies of saccade programming. In Ana Fred, Joaquim Filipe, and Hugo Gamboa, editors, *Biomedical Engineering Systems and Technologies*, volume 52 of *Communications in Computer and Information Science*, pages 187–199. Springer Berlin Heidelberg, 2010. 35

[73] Kenneth Holmqvist, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. *Eye Tracking. A comprehensive guide to methods and measures*. Oxford University Press, 2011. 23

[74] C. Hou, J. Wang, Y. Wu, and D. Yi. Local linear transformation embedding. *Neurocomputing*, 72:2368–2378, 2009. 98

[75] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 8, 16

[76] D.C. Howell. *Fundamental statistics for the behavioral sciences, Seventh Edition*. Linda Schreiber - Jon-David Hague, 2010. 29

[77] A. Hwang and M. Pomplun. A model of top-down control of attention during visual search in real-world scenes. *Journal of vision*, 8(6), 2008. 32

[78] Instagram. Year in review: 2011 in numbers. http://blog.instagram.com/post/15086846976/year-in-review-2011-in-numbers, 2012. 1

[79] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011. 3, 66, 67, 70, 71, 72, 73, 74

[80] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, Aug 2005. 16

[81] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Neural Information Processing Systems*, 2005. 8

[82] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009. 54

[83] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10:161–169, 2001. 12

[84] L. Itti, C. Koch, and E. Niebur. A model for saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20:1254–1259, 1998. 6, 9, 14, 16, 56, 66, 72, 112

[85] H. Jarodzka, K. Holmqvist, and M. Nystr. A vector-based, multidimensional scanpath similarity measure. In *Symposium on Eye-Tracking Research Applications*, pages 211–218, 2010. 47

[86] M.I Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorihtm. *Neural Computation*, 6:181–214, 1994. 81

[87] S. Josephson and M. E. Holmes. Attention to repeated images on the world-wide web: Another look at scanpath theory. *Behavior Research Methods, Instruments & Computers*, 34(4):539–548, 2002. 45

[88] T. Jost, N. Ouerhani, R. von Wartburg, R. Mauri, and H. Haugli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100:107–123, 2005. 49

[89] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *Journal of Vision*, 11:1–20, 2011. 25, 32, 83

[90] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixation. Technical report, MIT, 2012. 32

[91] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. Technical report, MIT (CSAIL-TR-2012-001), 2012. 12

[92] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where people look. In *ICCV*, 2009. 16, 32, 56, 76, 79, 116

[93] C. Kervrann and J. Boulanger. Local adaptivity to variable smoothness for examplar-based image denoising and representation. *International Journal of Computer Vision*, 79:45–69, 2008. 96

[94] D. Khaustova, J. Founier, E. Wyckens, and O. Le Meur. How visual attention is modified by disparities and textures changes? In *SPIE Human Vision and Electronic Imaging*, 2013. 32

[95] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, USA, December 2012. 66

[96] W. Kienzle, M.O. Franz, B. Scholkopf, and F.A. Wichmann. Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, 9:1–15, 2009. 32

[97] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985. 6

[98] N. Komodakis and G. Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Trans. On Image Processing*, 16(11):2649 – 2661, 2007. 103, 104, 107, 108

[99] G. Kootstra, B. de Boer, and L.R.B. Schomaler. Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation*, 3(1):223–240, 2011. 32

[100] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006. 70

[101] O. Le Meur. *Attention sélective en visualisation d'images fixes et animées affichées sur écran: modèles et évaluation de performances - applications*. PhD thesis, University of Nantes, 2005. 25

[102] O. Le Meur. Predicting saliency using two contextual priors: the dominant depth and the horizon line. In *ICME*, 2011. 12, 13, 116

[103] O. Le Meur. Robustness and repeatability of saliency models subjected to visual degradations. In *ICIP*, pages 3285 – 3288, 2011. 14, 16

[104] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Method*, 1:1–16, 2012. 40, 69, 95, 106, 111

[105] O. Le Meur, T. Baccino, and A. Roumy. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *ACM Multimedia*, 2011. 80, 91

[106] O. Le Meur, X. Castellan, P. Le Callet, and D. Barba. Efficient saliency-based repurposing method. In *ICIP*, 2006. 2

[107] O. Le Meur, M. Ebdelli, and C. Guillemot. Hierarchical super-resolution-based inpainting. *IEEE Transactions on Image Processing*, 22(10):3779–3790, August 2013. 102, 107

[108] O. Le Meur and P. Le Callet. What we see is most likely to be what matters: visual attention and applications. In *ICIP*, pages 3085–3088, 2009. 6

[109] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47:2483–2498, 2007. 19, 26, 111

[110] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model the bottom-up visual attention. *IEEE Trans. On PAMI*, 28(5):802–817, May 2006. 2, 9, 10, 14, 16, 32, 38, 40, 49, 56, 82, 111

[111] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba. Do video coding impairments disturb the visual attention deployment? *Signal Processing: Image Communication*, 25(8):597–609, 2010. 61

[112] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba. Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7):547–558, 2010. 62

[113] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Process. Syst. (NIPS)*, 2000. 98

[114] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2001. 98

[115] B. Lemaire, A. Guérin-Dugué, T. Baccino, M. Chanceaux, and L. Pasqualotti. A cognitive computational model of eye movements investigating visual strategies on textual material. In *Annual Conference of the Cognitive Science Society*, 2011. 47

[116] R.V. Lenth. Some practical guidelines for effective sample size determination. *The American Statistician*, 55:187–193, 2001. 29

[117] P. L. Lester. *Visual Communication: Images with Messages, 6th ed.* Cengage Learning, Wadsworth series in mass communication and journalism, 2012. 61

[118] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, 6:707–710, 1966. 45

[119] A. Levin. Blind motion deblurring using image statistics. In *NIPS*, 2006. 78

[120] J. Li, M.D. Levine, X. An, X. Xu, and H. He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE trans. on Pattern and Machine Intelligence*, 35(4):996–1010, 2012. 115

[121] Jian Li, Martin D. Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 996–1010, 2013. 32

[122] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *ICIP*, volume 1, pages 900–903, 2002. 77

[123] H. Liu and I. Heynderickx. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *ICIP*, 2009. 32

[124] C. Louchet and L. Moisan. Total variation as a local filter. *SIAM J. Imaging Sciences*, 4(2):651–694, 2011. 95

[125] Y. Luo and X. Tang. Photo and video quality evaluation: focussing on the subject. In *ECCV*, pages 386–399, 2008. 78, 80, 83

[126] M. Mancas and O. Le Meur. Memorability of natural scene: the role of attention. In *ICIP*, 2013. 3, 32

[127] S.K. Mannan, K.H. Ruddock, and D.S. Wooding. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2d images. *Spatial Vision*, 9:363–386, 1995. 25, 46

[128] S.K. Mannan, K.H. Ruddock, and D.S. Wooding. Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26(8):1059–1072, 1997. 25, 46

[129] B. Musel, A. Chauvin, N. Guyader, S. Chokrom, and C. Peyrin. Is coarse-to-fine strategy sensitive to normal aging? *Plos One*, 7:1–6, 2012. 28

[130] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *ICIP*, 2007. 61

[131] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing, Special Issue On Visual Media Quality Assessment*, 3(2):253 – 265, 2009. 63, 64

[132] R. Nisbett. *The geography of thought: how Asians and Westerners think differently... and why.* New York: Free Press, 2003. 27

[133] A. Nuthmann, T. J. Smith, R. Engbert, and J. M. Henderson. CRISP: A Computational Model of Fixation Durations in Scene Viewing. *Psychological Review*, 117(2):382–405, April 2010. 115

[134] M. Nystrom and K. Holmqvist. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1):188–204, 2010. 23

[135] A. Oliva, M.L. Mack, M. Shrestha, and A. Peeper. Identifying the perceptual dimensions of visual complexity of scenes. In *26th annual meeting of the Cognitive Science Society Meeting*, 2004. 80

[136] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 70

[137] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-down control of visual attention in object detec- tion. In *IEEE ICIP*, 2003. 7, 8

[138] D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002. 26, 30

[139] R. J. Peters and L. Itti. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception*, pages 1–21, 2008. 54

[140] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005. 53

[141] Y. Pritch, E. Kav-Venaki, and S. Peleg. Shift-map image editing. In *ICCV'09*, pages 151–158, Kyoto, Sept 2009. 107, 108

[142] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000. 45

[143] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack. Gaffe: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing*, 17:564–573, 2008. 49

[144] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. In *ECCV*, 2010. 32

[145] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psyhological bulletin*, 124(3):372–422, November 1998. 33, 75, 115

[146] J. Redi, H. Liu, R. Zumino, and I. Heynderickx. Interations of visual attention and quality perception. In *SPIE HVEI*, volume 7865, 2011. 32

[147] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. A study of parameters affecting visual saliency assessment. In *6th International Symposium on Attention in Cognitive Systems*, 2013. 115

[148] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642658, 2013. 71

[149] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of Vision*, 7(2), March 2007. 38, 75, 80, 82, 84

[150] M.G. Ross and A. Oliva. Estimating perception of scene layout properties from global image features. *Journal Of Vision*, 10(1), Januray 2010. 80

[151] C. Rother, L. Bordeaux, Y. Hamadi, and A. Black. Autocollage. In *in ACM Transactions on Graphics (SIGGRAPH)*, 2006. 83

[152] D. Rouse and S. Hemami. Natural image utility assessment using image contours. In *ICIP*, pages 2217 – 2220, 2009. 62

[153] S. Salvucci and J. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Eye Tracking Research and Application (ETRA)*, page 7178, 2000. 23

[154] L.K. Saul and S.T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003. 97

[155] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 70

[156] F. Shic, K. Chawarska, and B. Scassellati. The incomplete fixation measure. In ACM, editor, *Eye Tracking Research and Application (ETRA)*, pages 111–114, 2008. 23, 24

[157] J. Simola, J. Salojarvi, and I. Kojo. Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9(4):237–251, 2008. 44

[158] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computatinal visual attention model. In *ACM Multimedia*, pages 541–544, 2009. 83

[159] V. N. Tam, B. Ni, H. Liu, W. Xia, J. Luo, M. Kankanhalli, and S. Yan. Image re-attentionizing. *IEEE Transactions on Multimedia*, 15(8):1910–1919, 2013. 118

[160] B. W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7, 2007. 42

[161] B.W. Tatler, R. J. Baddeley, and I.D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45:643–659, 2005. 26, 27, 30, 52, 53, 69

[162] A. Toet. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33:2131–2146, 2011. 49

[163] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Pattern Analysis and Machine Intelligence*, 24(9):1226–1238, 2002. 80

[164] A. Torralba and A. Oliva. Statistics of natural image catagories. *network*, 14:391–421, 2003. 84

[165] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766–786, 2006. 36, 52, 56, 62, 76, 115

[166] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. 6, 13

[167] P. Tseng, R. Carmi, I. G. M. Cameron, D.P. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7:4):1–16, July 2009. 39

[168] G. Underwood and T. Foulsham. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly journal of experimental psychology*, 59(11):1931–1949, 2006. 77

[169] P.J.A. Unema, S. Pannasch, M. Joos, and B.M. Velichkovsky. Time course of information processing during scene perception: the relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12(3):473–494, 2005. 33, 34

[170] D. Van De Ville and M. Kocher. Sure-based non-local means. *IEEE Signal Processing Letters*, 16(11):973–976, 2009. 96

[171] I. Van Der Linde, U. Rajashekar, A.C. Bovik, and L.K. Cormack. Doves: a database of visual eye movements. *Spatial Vision*, 22(2):161–177, 2009. 32

[172] B. M. Velichkovsky. Heterarchy of cognition: The depths and the highs of a framework for memory research. *Memory*, 10:405419, 2002. 24

[173] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 77

[174] P. Viviani. Eye movements in visual search: cognitive, perceptual and motor control aspects. *Reviews of Oculomotor Research*, pages 353–393, 1990. 44

[175] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974. 46

[176] J. Wang, D.M. Chandler, and P. Le Callet. Quantifying the relationship between visual salience and visual importance. In *SPIE HVEI*, volume 7527, pages 17–21, 2010. 32

[177] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao. Simulating human saccadic scanpaths on natural images. In *CVPR*, 2011. 112, 113

[178] J. Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 32:111–127, 1999. 92, 93

[179] M. Wexler and N. Ouarti. Depth affects where we look. *Current Biology*, 18(23):1872–1876, 2008. 115

[180] Y. Wexler, E. Shechtman, and M. Irani. Space-time video completion. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2004. 91, 96

[181] S. Winkler and R. Subramanian. Overview of eye tracking datasets. In *5th international Workshop on Quality of Multimedia Experience (QoMEX)*, 2013. 30, 32

[182] A. Wong and J. Orchard. A nonlocal-means approach to exemplar-based inpainting. In *IEEE Int. Conf. Image Processing (ICIP)*, pages 2600–2603, 2006. 96

[183] L.K Wong and K.L. Low. Saliency retargeting: an approach to enhance image aesthetics. In *WACV*, pages 73–80, 2011. 118

[184] Z. Xu and J. Sun. Image inpainting by patch propagation using patch sparsity. *IEEE Trans. on Image Processing*, 19(5):1153–1165, 2010. 91

[185] P. Yanilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *4th Annu ACM/SIGACT-SIAM Symp; discrete Algorithms*, pages 311–321, 1993. 95

[186] A.L. Yarbus. *Eye movements and vision.* Plenum Press: New York, 1967. 26, 27, 44

[187] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *J.S. Yedidia, W.T. Freeman, and Y. Weiss*, 51:2282–2312, 2005. 105

[188] C-G Yeh, Y-C Ho, B.A. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *ACM Multimedia*, 2010. 83

[189] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for salience using natural statistics. *Journal of Vision*, 8(7):1–20, 2008. 8, 55

[190] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal Of Vision*, 11(3):1–9, 2011. 115